Independent Natural Policy Gradient Methods for Potential Games: Finite-time Global Convergence with Entropy Regularization

Shicong Cen*†
Carnegie Mellon University

Fan Chen*‡
Peking University

Yuejie Chi†
Carnegie Mellon University

April 11, 2022

Abstract

A major challenge in multi-agent systems is that the system complexity grows dramatically with the number of agents as well as the size of their action spaces, which is typical in real world scenarios such as autonomous vehicles, robotic teams, network routing, etc. It is hence in imminent need to design decentralized or independent algorithms where the update of each agent is only based on their local observations without the need of introducing complex communication/coordination mechanisms.

In this work, we study the finite-time convergence of independent entropy-regularized natural policy gradient (NPG) methods for potential games, where the difference in an agent's utility function due to unilateral deviation matches exactly that of a common potential function. The proposed entropy-regularized NPG method enables each agent to deploy symmetric, decentralized, and multiplicative updates according to its own payoff. We show that the proposed method converges to the quantal response equilibrium (QRE)—the equilibrium to the entropy-regularized game—at a sublinear rate, which is independent of the size of the action space and grows at most sublinearly with the number of agents. Appealingly, the convergence rate further becomes *independent* with the number of agents for the important special case of identical-interest games, leading to the first method that converges at a *dimension-free* rate. Our approach can be used as a smoothing technique to find an approximate Nash equilibrium (NE) of the unregularized problem without assuming that stationary policies are isolated.

1 Introduction

Reinforcement learning (RL) has garnered a growing amount of interest in recent years, due to its excellent empirical performance in a wide variety of applications, such as Go [SHM+16], motor control [LFDA16], chip design [MGY+21], and so on. There is a growing interest to applying RL techniques such as Q-learning and policy gradient methods to the setting with multiple agents, i.e., multi-agent reinforcement learning (MARL) problems.

While it seems appealing to apply single-agent RL methods to each agent in a multi-agent system in a straightforward fashion, this approach neglects non-stationarity of the environment due to the presence of other agents, and thus lacks theoretical support in general. The complication has given rise to the paradigm of *centralized training with decentralized execution* (CTDE) [LWT+17], where the policies are first obtained through training with a centralized controller with access to all agents' observations and then disseminated to each agent for execution. However, this approach falls short of adapting to changes in the environment without retraining and raises privacy concerns as well. It is hence of great interest to understand and design more versatile *independent learning* algorithms that only depend on the agents' local observations, require minimal coordination between agents, and provably converge.

---

*The first two authors contributed equally.

†S. Cen and Y. Chi are with the Department of Electrical and Computer Engineering, Carnegie Mellon University; emails: {shicongc,yuejiec}@andrew.cmu.edu.

‡F. Chen is with the Department of Mathematics at Peking University; email: chern@pku.edu.cn.

In this work, we focus on independent learning algorithms for potential games [MS96b], an important class of games that admit a potential function to capture the differences in each agent's utility function induced by unilateral deviations. Despite decades-long research, however, the computational underpinnings of such problems are still far from mature, especially when it comes to finding the Nash equilibrium (NE) of potential games in a decentralized manner. While several recent works have made significant breakthroughs by achieving logarithmic regrets with independent learning dynamics [DFG21, ADF+21], these results only guarantee convergence to coarse correlated equilibrium or correlated equilibrium, which are arguably much weaker equilibrium concepts than NE and hence do not lead to an approximate NE solution.

1.1 Our contributions

We seek to find the quantal response equilibrium (QRE) [MP95], the prototypical extension of NE for games with bounded rationality [Sel89], where each agent runs independent natural policy gradient (NPG) methods [Kak01] involving symmetric, decentralized, and multiplicative updates according to its own payoff. This amounts to solving a potential game with entropy regularization, whose algorithmic role has been studied in the setting of single-agent RL [MXSS20, CCC+21] as well as two-player zero-sum games [CWC21], but yet to be explored in more general settings. Our contributions are summarized below.

- **Finite-time global convergence of independent entropy-regularized NPG methods.** We show that independent entropy-regularized NPG methods provably converge to the QRE of a potential game, and it takes no more than

$$O\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}}{\tau^2 \epsilon^2}\right)$$

  iterations to find an ϵ-optimal QRE (to be defined precisely). Here, N stands for the number of agents, $\tau > 0$ the entropy regularization parameter, and $\Phi_{\max} > 0$ the maximum value of the potential function.

- **Finite-time global convergence to ϵ-NE without isolation assumption.** By setting the entropy regularization parameter τ sufficiently small, the result translates to finding an approximate NE with non-asymptotic convergence guarantees, thereby obviating the additional assumption in prior literature [FMOP21, PPP17, ZMD+22] that requires the set of stationary policies to be isolated. Specifically, it takes no more than

$$\widetilde{O}\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}}{\epsilon^4}\right)$$

  iterations to find an ϵ-NE for the unregularized potential game, where $\widetilde{O}$ hides logarithmic dependencies.

These rates give the first set of iteration complexities—to the best of our knowledge—that are independent of the size of the action spaces, up to logarithmic factors. In addition, the iteration complexities exhibit a sublinear dependency with the number of agents, outperforming existing NE-finding algorithms whose complexities depend at least linearly with the number of agents. Even more appealingly, when interpreting our convergence rates for the important special case of identical-interest games with bounded payoffs [MS96a], they further become independent with the number of agents, leading to the *first* method that achieves a *dimension-free* convergence rate of $\widetilde{O}\left(1/\epsilon^4\right)$ to find an ϵ-NE.

1.2 Related works

We review some related works, focusing on the theoretical advances on policy gradient methods and independent learning in games.

Global convergence of policy gradient methods Only recently theoretical understandings on the global convergence of policy gradient (PG) methods have emerged, mostly in the single-agent setting, including but not limited to [FGKM18, AKLM20, MXSS20, CCC+21, BR19, LZBY20, LWC+21, MXD+20,

WCYW19, Xia22]. In addition, several works developed finite-time guarantees of independent PG methods for zero-sum two-player Markov games [DFG20, WLZL21, ZTLD21, CWC21] in the competitive MARL setting. A recent line work has been successful in extending PG methods with direct parameterization to Markov potential games [ZRL21, LOPP21, DWZJ22], by validating the gradient dominance property [AKLM20]. In addition, [ZMD$^+$22] studies the finite-time convergence rate of PG methods with softmax parameterization for Markov potential games. Given that NPG methods often have better finite-time convergence rates than vanilla PG methods in the single-agent setting, our work focuses on the understanding of NPG methods for potential games.

Fast convergence of natural policy gradient methods with entropy regularization Entropy regularization as a de facto trick to promote exploration in RL [HZAL18] and has been shown to provably accelerate convergence of policy gradient methods for single-agent RL [MXSS20, CCC$^+$21, CWC21, ZCH$^+$21, Lan21]. In particular, combining entropy regularization with NPG methods leads to fast linear convergence at a desirable dimension-free rate [CCC$^+$21, ZCH$^+$21, Lan21], which continues to hold in two-player zero-sum games [CWC21]. Extending such results to potential games, however, is non-trivial, due to the non-uniqueness of NE even with regularization. [FMOP21, PPP17, ZMD$^+$22] established asymptotic convergence of independent NPG methods for Markov potential games with an additional assumption that requires the set of stationary policies to be isolated. [HCM17] demonstrated asymptotic convergence of NPG with diminishing step sizes for potential games in the bandit feedback setting. In addition, [ZMD$^+$22] proposed to use a log-barrier regularization along with NPG to sidestep the isolation assumption and achieved the same iteration complexity as that of PG methods with direct parameterization. In contrast, we consider NPG with entropy regularization, which achieves a convergence rate that has better dependencies with the size of the action spaces and the number of agents.

Independent learning in general-sum games Considerable progress has been made towards understanding independent learning dynamics in general-sum games [DFG21, ADF$^+$21] and general-sum Markov games (also known as stochastic games) [SMB21, JLWY21, MB22] by establishing non-asymptotic convergence to correlated equilibrium and coarse correlated equilibrium. However, such successes do not directly extend to potential games where NE is of interest. Specialized analysis for potential games is thus needed as finding approximate NE in a two-player game can be PPAD-hard even with full information [Das13]. Notably, [SMB21] proposed a method for solving Markov potential games, yet the method therein requires policies to be updated one agent at a time and therefore does not fall under the independent learning paradigm.

1.3 Notation and paper organization

We use $\Delta(\mathcal{A})$ to denote the probability simplex over the set $\mathcal{A}$. For a vector $\mathbf{a} \in \mathcal{A}^N$, we use $a_i \in \mathcal{A}$ and $a_{-i} \in \mathcal{A}^{N-1}$ to denote the entry with index i and all the rest entries as a vector, respectively. The application of scalar functions such as exp and log to vectors are defined in an entry-wise fashion. Let $\mathbf{1}$ be the all-one vector, and $[N] = \{1, \ldots, N\}$. Given two distributions π_1 and π_1' over $\mathcal{A}$, the Kullback-Leibler (KL) divergence from π_1' to π_1 is defined by $\mathsf{KL}\big(\pi_1 \,\|\, \pi_1'\big) = \sum_{a \in \mathcal{A}} \pi_1(a)(\log \pi_1(a) - \log \pi_1'(a))$. Note that KL divergence is additive for product distributions in the sense that $\mathsf{KL}\big(\pi \,\|\, \pi'\big) = \sum_{i \in [N]} \mathsf{KL}\big(\pi_i \,\|\, \pi_i'\big)$ for $\pi = \pi_1 \times \cdots \times \pi_N \in \Delta(\mathcal{A})^N$ and $\pi' = \pi_1' \times \cdots \times \pi_N' \in \Delta(\mathcal{A})^N$. We denote Jeffrey divergence [Jef98] by $J(\pi, \pi') = \mathsf{KL}\big(\pi \,\|\, \pi'\big) + \mathsf{KL}\big(\pi' \,\|\, \pi\big)$, which is the symmetric version of the KL divergence.

The rest of this paper is organized as follows. Section 2 presents the backgrounds of the potential game setup. Section 3 introduces independent NPG methods and presents the finite-time global convergence guarantees. Section 4 provides an outline to the analysis and the rest of the proofs are deferred to the appendix. Section 5 presents numerical results to verify the theoretical findings. Finally, we conclude in Section 6.

2 Potential Games with Entropy Regularization

In this section, we introduce the basics of potential games, as well as the incorporation of entropy regularization into its formulation.

2.1 Potential games

A strategic game $\mathcal{G} = \{N, \mathcal{A}, \{u_i\}_{i \in [N]}\}$ consists of N agents each with an individual utility or payoff function

$$u_i : \mathcal{A}^N \to [0, 1], \qquad i \in [N],$$

where $\mathcal{A}$ is, without loss of generality, a finite action space shared by all agents. The policy or mixed strategy of agent i is denoted by $\pi_i \in \Delta(\mathcal{A})$, which is a distribution over the action space $\mathcal{A}$. By an abuse of notation, let $u_i(\pi)$ denote agent i's expected utility function under the joint policy $\pi = \pi_1 \times \cdots \times \pi_N \in \Delta(\mathcal{A})^N$, i.e.,

$$u_i(\pi) = \mathbb{E}_{a_i \sim \pi_i, \forall i \in [N]} [u_i(\boldsymbol{a})],$$

where we denote the action profile $(a_1, \cdots, a_N)$ of all agents by $\boldsymbol{a} \in \mathcal{A}^N$. We shall often instead write $\boldsymbol{a} = (a_i, a_{-i})$ where $a_{-i} = \{a_j\}_{j \neq i}$ collects the actions of all agents but i; similarly, we write $\pi = (\pi_i, \pi_{-i})$, where $\pi_{-i} = \{\pi_j\}_{j \neq i}$ collects the policies of all agents but i.

The game $\mathcal{G}$ is said to be a potential game if there exists a potential function $\Phi : \mathcal{A}^N \to \mathbb{R}$ such that

$$u_i(a_i, a_{-i}) - u_i(a'_i, a_{-i}) = \Phi(a_i, a_{-i}) - \Phi(a'_i, a_{-i})$$

for any $a_i, a'_i \in \mathcal{A}$, $a_{-i} \in \mathcal{A}^{N-1}$ and $i \in [N]$. We assume that

$$0 \leq \Phi(\boldsymbol{a}) \leq \Phi_{\max}, \qquad \forall \boldsymbol{a} \in \mathcal{A}^N, \tag{1}$$

where $\Phi_{\max}$ upper bounds the potential function. An important special case of the potential game is when all the agents share the same utility function, known as the identical-interest game [MS96a]. It is straightforward to see that for an identical-interest game, we can set $\Phi = u_i$ for all $i \in [N]$, and therefore $\Phi_{\max} = 1$ due to the fact that the individual payoff is bounded in $[0, 1]$.

By linearity of expectation, we have

$$u_i(\pi_i, \pi_{-i}) - u_i(\pi'_i, \pi_{-i}) = \Phi(\pi_i, \pi_{-i}) - \Phi(\pi'_i, \pi_{-i}),$$

where, again with slight abuse of notation, we denote

$$\Phi(\pi) = \mathbb{E}_{\boldsymbol{a} \sim \pi} [\Phi(\boldsymbol{a})] = \mathbb{E}_{a_i \sim \pi_i, \forall i \in [N]} [\Phi(\boldsymbol{a})],$$

for any $\pi_i, \pi'_i \in \Delta(\mathcal{A})$, $\pi_{-i} \in \Delta(\mathcal{A})^{N-1}$ and $i \in [N]$.

Nash equilibrium We now introduce the important notion of *Nash equilibrium* in a potential game.

Definition 2.1 (Nash equilibrium). *A joint policy $\pi^\star$ is called a Nash equilibrium (NE) when it holds that*

$$u_i(\pi_i, \pi_{-i}) \geq u_i(\pi'_i, \pi_{-i}), \quad \forall \pi'_i \in \Delta(\mathcal{A}), \forall i \in [N].$$

In other words, every agent cannot improve its utility function by deviating from the current policy. It is known that there exists at least one NE in a strategic game with finite agents and actions [Nas51]. It follows immediately that the policy or strategy profile maximizing Φ in a potential game is an NE.

Marginalized utility Before continuing, let us introduce an important quantity called the marginalized utility $r_i^\pi : \mathcal{A} \to \mathbb{R}$:

$$r_i^\pi(a) = \mathbb{E}_{a_{-i} \sim \pi_{-i}} [u_i(a, a_{-i})], \tag{2}$$

which can be viewed as the "single-agent" payoff or reward function when the policies of other agents are fixed. It is immediate to see that the utility function u_i can be written as

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

Here and throughout this paper, we shall often abuse the notation to treat π, π_i and $r_i^{(t)}$ as vectors.

2.2 Entropy-regularized potential games

The *quantal response equilibrium* (QRE) is proposed by McKelvey and Palfrey [MP95] as a seminal extension to the Nash equilibrium, which enables players to combat randomness in payoffs. A QRE or logit equilibrium $\pi_\tau^\star = \pi_{\tau,1}^\star \times \cdots \times \pi_{\tau,N}^\star$ necessitates every agent to maximize its own utility function with entropy regularization [MS16], i.e.,

$$u_{i,\tau}(\pi_{\tau,i}^\star, \pi_{\tau,-i}^\star) \geq u_{i,\tau}(\pi_i', \pi_{\tau,-i}^\star), \quad \forall \pi_i' \in \Delta(\mathcal{A}),$$

where the entropy-regularized individual utility function is given by

$$u_{i,\tau}(\pi) = u_i(\pi) + \tau \mathcal{H}(\pi_i).$$

Here, $\pi = \pi_1 \times \cdots \times \pi_N$, $\tau > 0$ is the regularization parameter, and $\mathcal{H}(\pi_i) = -\sum_{a \in \mathcal{A}} \pi_i(a|s) \log \pi_i(a|s)$ is the Shannon entropy of the policy π employed by agent i. By introducing the regularized potential function

$$\Phi_\tau(\pi) = \Phi(\pi) + \tau \mathcal{H}(\pi) := \Phi(\pi) + \tau \sum_{i \in [N]} \mathcal{H}(\pi_i),$$

it is easy to verify

$$u_{i,\tau}(\pi_i, \pi_{-i}) - u_{i,\tau}(\pi_i', \pi_{-i}) = \Phi_\tau(\pi_i, \pi_{-i}) - \Phi_\tau(\pi_i', \pi_{-i}).$$

for any $\pi_i, \pi_i' \in \Delta(\mathcal{A}), \pi_{-i} \in \Delta(\mathcal{A})^{N-1}$ and $i \in [N]$, as long as the unregularized game is a potential game.

Fixed-point characterization of QRE An equivalent interpretation of QRE is to let each agent assign the probability mass in its policy according to every action's utility in a bounded rationality fashion [Sel89]:

$$\pi_{\tau,i}^\star(a) \propto \exp\left(r_i^{\pi_\tau^\star}(a)/\tau\right), \quad \forall i \in [N], \tag{3}$$

where $r_i^{\pi_\tau^\star}$ is the marginalized utility of $\pi_\tau^\star$ defined in (2). Note that the above relation defines a fixed-point equation of $\pi_\tau^\star$.

3 Finite-Time Global Convergence of Independent Natural Policy Gradient Methods

A popular approach in the game theory literature to find an NE of a potential game is for each agent to switch to the best response policy, one at a time. This approach converges to an NE in finite iterations, yet the update sequence needs to be scheduled in a centralized manner that is not often possible, and the number of iterations can be exponentially large [Dur18, SV08]. Therefore, it is greatly desirable to design independent update rules, where each agent updates simultaneously without observing the payoffs of other agents, that achieves faster convergence. In this section, we answer this call by developing the independent natural policy gradient method to solve (entropy-regularized) potential games with finite-time global convergence guarantees.

3.1 Independent natural policy gradient method

In policy optimization, it is common practice to parameterize the policy class in a way that obviates the need for tackling probability simplex constraint. We consider the standard softmax parameterization, where every agent i generates its own policy π_{θ_i} parameterized with $\theta_i \in \mathbb{R}^{|\mathcal{A}|}$ through the softmax transform:

$$\pi_{\theta_i}(a) = \frac{\exp(\theta_i(a))}{\sum_{a \in \mathcal{A}} \exp(\theta_i(a))}.$$

Every agent i evaluates and updates its policy independently using the *natural policy gradient* (NPG) method [Kak01]:

$$\theta_i \leftarrow \theta_i + \eta (\mathcal{F}^{\theta_i})^\dagger \nabla_{\theta_i} u_{i,\tau}(\pi), \tag{4}$$

where $\mathcal{F}^{\theta_i}$ denotes the Fisher information matrix defined as

$$\mathcal{F}^{\theta_i} = \mathbb{E}_{a \sim \pi_{\theta_i}(\cdot)} \left[(\nabla_{\theta_i} \log \pi_{\theta_i}(a))(\nabla_{\theta_i} \log \pi_{\theta_i}(a))^\top \right],$$

and $\eta > 0$ is the learning rate. Moreover, the gradient $\nabla_{\theta_i} u_{i,\tau}(\pi)$ can be expressed as

$$\nabla_{\theta_i} u_{i,\tau}(\pi) = r_i^\pi - \tau \log \pi_i - \tau \mathbf{1}.$$

It turns out that with some algebra, the NPG update rule (4) can be equivalently rewritten with respect to the policies in use [CCC+21]:

$$\pi_i^{(t+1)}(a) \propto \pi_i^{(t)}(a)^{1-\eta\tau} \exp(\eta r_i^{(t)}(a)), \tag{5}$$

where $\pi_i^{(t)}$ denotes agent i's policy in the t-th iteration, and $r_i^{(t)} := r_i^{\pi^{(t)}}$ denotes the marginalized utility of $\pi^{(t)}$ (cf. (2)).

To better understand the update rule (5) as well as prepare for follow-up analysis, we introduce $\pi_i^{\star(t)}$ to denote agent i's best-response policy in the t-th iteration, which is the policy that obeys

$$u_{i,\tau}(\pi_i^{\star(t)}, \pi_{-i}^{(t)}) = \max_{\pi_i'} u_{i,\tau}(\pi_i', \pi_{-i}^{(t)}). \tag{6}$$

It is easily seen that

$$\pi_i^{\star(t)}(a) \propto \exp(r_i^{(t)}(a)/\tau). \tag{7}$$

Therefore, the updated policy in (5) can be regarded as a multiplicative combination of the current policy $\pi_i^{(t)}$ and the best-response policy $\pi_i^{\star(t)}$, where the weight is controlled by the learning rate η.

3.2 Finite-time global convergence

We are now ready to present our main theorem concerning the finite-time global convergence of independent NPG for solving entropy-regularized potential games. We introduce

$$\texttt{NE-gap}(\pi) = \max_{i \in [N], \pi_i' \in \Delta(\mathcal{A})} \left[u_i(\pi_i', \pi_{-i}) - u_i(\pi_i, \pi_{-i}) \right]$$

and

$$\texttt{QRE-gap}_\tau(\pi) = \max_{i \in [N], \pi_i' \in \Delta(\mathcal{A})} \left[u_{i,\tau}(\pi_i', \pi_{-i}) - u_{i,\tau}(\pi_i, \pi_{-i}) \right]$$

to characterize how close the joint policy π is to an equilibrium. A joint policy π is said to be an ϵ-QRE (resp. ϵ-NE) when $\texttt{QRE-gap}(\pi) \leq \epsilon$ (resp. $\texttt{NE-gap}(\pi) \leq \epsilon$). For notational simplicity, we denote

$$\Phi_\tau^{(t)} := \Phi_\tau(\pi^{(t)}), \qquad \texttt{QRE-gap}_\tau^{(t)} := \texttt{QRE-gap}_\tau(\pi^{(t)}), \quad \text{and} \quad \texttt{NE-gap}^{(t)} := \texttt{NE-gap}(\pi^{(t)}).$$

Our main theorem is as follows, whose proof is deferred to Section 4.

Theorem 3.1. *Suppose that the learning rate η satisfies $\eta \leq \frac{1}{2(\min\{\sqrt{N}, 2\Phi_{\max}\} + \tau)}$, then for independent NPG updates (5), it holds that*

$$\frac{1}{T} \sum_{t=1}^T \texttt{QRE-gap}_\tau^{(t)} \leq \frac{2}{\eta\tau T} \left(\tau \left\| \log \pi^{(0)} - \log \pi^{\star(0)} \right\|_\infty + \sqrt{2\eta T (\Phi_\tau^{(T)} - \Phi_\tau^{(0)})} \right).$$

Theorem 3.1 suggests that the average iterate of independent NPG converges to an ϵ-QRE at a sublinear rate when we initialize it via uniform policies, as indicated in the following corollary. The proof can be found in Appendix B.3.

Corollary 3.2. *Assume the independent NPG method is initialized with uniform policies at all agents. Setting the learning rate $\eta = 1/(2(\min\{\sqrt{N}, 2\Phi_{\max}\} + \tau))$ and $\tau = O(1)$, then independent NPG updates ensure that $\frac{1}{T} \sum_{t=1}^T \texttt{QRE-gap}_\tau^{(t)} \leq \epsilon$ with at most*

$$T = O\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\} \Phi_{\max}}{\tau^2 \epsilon^2} \right)$$

iterations.

Finding approximate NEs. It is possible to leverage the entropy-regularized potential game to find an approximate NE by setting the regularization parameter sufficiently small. Note that

$$
\begin{aligned}
\texttt{NE-gap}(\pi) &= \max_{i \in [N], \pi_i' \in \Delta(\mathcal{A})} \left[u_i(\pi_i', \pi_{-i}) - u_i(\pi_i, \pi_{-i}) \right] \\
&\leq \max_{i \in [N], \pi_i' \in \Delta(\mathcal{A})} \left[u_{i,\tau}(\pi_i', \pi_{-i}) - u_{i,\tau}(\pi_i, \pi_{-i}) \right] + \max_{i \in [N], \pi_i' \in \Delta(\mathcal{A})} \left[-\tau \mathcal{H}(\pi_i') + \tau \mathcal{H}(\pi_i) \right] \\
&\leq \texttt{QRE-gap}_\tau(\pi) + \tau \log |\mathcal{A}|.
\end{aligned}
$$

Therefore, by setting the entropy regularization at

$$
\tau = \frac{\epsilon}{2 \log |\mathcal{A}|},
$$

with at most

$$
T = \widetilde{O} \left(\frac{\min\{\sqrt{N}, 2\Phi_{\max}\}\Phi_{\max}}{\epsilon^4} \right)
$$

iterations, we can ensure $\frac{1}{T} \sum_{t=1}^{T} \texttt{NE-gap}^{(t)} \leq \epsilon$.

Comparisons with prior art Importantly, our iteration complexities do not depend on the size of the action space (up to logarithmic factors), which is in sharp contrast to existing analyses of potential games using other policy gradient approaches, such as direct PG [ZRL21, LOPP21, DWZJ22] and NPG with log-barrier regularization [ZMD$^+$22], where the iteration complexity scales as $\widetilde{O}\left(\frac{N|\mathcal{A}|\Phi_{\max}}{\epsilon^2}\right)$ to find an ϵ-approximate NE. In comparison, while our rate $\widetilde{O}\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}}{\epsilon^4}\right)$ is worse in terms of ϵ, it is almost independent of the size $|\mathcal{A}|$ of the action space, as well as exhibits only a sublinear dependency with the number of agents N, thus can be beneficial for problems with large action spaces and a large number of agents. Furthermore, for the special case of identical-interest games [MS96a] where $\Phi_{\max} = 1$, the convergence rate of our method simplifies to

$$
\widetilde{O}\left(\frac{1}{\epsilon^4}\right),
$$

which leads to the first method that achieves a dimension-free iteration complexity (up to a logarithmic factor) for finding an ϵ-NE without imposing any isolation assumptions.

4 Proof of Theorem 3.1

Before proceeding to the main proof, we first record two useful lemmas. The following elementary lemma is standard (see e.g., [CWC21, Lemma 3] and [CCC$^+$21]) and will be helpful in the analysis.

Lemma 4.1. *For any $\mu_1, \mu_2 \in \Delta(\mathcal{A})$ satisfying*

$$
\mu_1(a) \propto \exp(x_1(a)) \quad \text{and} \quad \mu_2(a) \propto \exp(x_2(a))
$$

for some $x_1, x_2 \in \mathbb{R}^{|\mathcal{A}|}$, we have

$$
\| \log \mu_1 - \log \mu_2 \|_\infty \leq 2 \| x_1 - x_2 \|_\infty. \tag{8}
$$

Another useful lemma connects the marginalized utility with the policy, as given below.

Lemma 4.2. *Given any $\pi, \pi' \in \Delta(\mathcal{A})^N$, the difference in the marginalized utility (cf. (2)) can be bounded by*

$$
\left\| r_i^\pi - r_i^{\pi'} \right\|_\infty \leq \sqrt{J(\pi, \pi')},
$$

where $J(\pi, \pi') = \mathsf{KL}(\pi \,\|\, \pi') + \mathsf{KL}(\pi' \,\|\, \pi)$ is the Jeffrey divergence.

Proof. See Appendix A. $\qquad\square$

4.1 Step 1: quantify the policy improvement

We start by the following key lemma that gives a lower bound of the improvement in terms of the regularized potential function $\Phi_\tau^{(t)}$.

Lemma 4.3. *The independent NPG update* (5) *guarantees that*

$$\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \left(\frac{1}{\eta} - \min\{\sqrt{N}, 2\Phi_{\max}\} - \tau \right) J\left(\pi^{(t+1)}, \pi^{(t)} \right).$$

Proof. See Appendix B. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Lemma 4.3 ensures the monotonic improvement of the regularized potential function Φ_τ when η is not too large. Specifically, setting $\eta \leq 1/(2(\min\{\sqrt{N}, 2\Phi_{\max}\} + \tau))$, we have

$$\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \frac{1}{2\eta} J\left(\pi^{(t+1)}, \pi^{(t)} \right),$$

which is guaranteed to be non-negative. Summing the above inequality over $t = 0, \cdots, T-1$ gives

$$\sum_{t=0}^{T-1} J\left(\pi^{(t+1)}, \pi^{(t)} \right) \leq 2\eta(\Phi_\tau^{(T)} - \Phi_\tau^{(0)}), \tag{9}$$

which controls the change of $\pi^{(t)}$ over time t via the size of the regularized potential function.

4.2 Step 2: introduce the auxiliary sequence

Motivated by [CWC21, CCC$^+$21], we introduce an auxiliary sequence $\{\xi_i^{(t)} \in \mathbb{R}^{|\mathcal{A}|}, \ i \in [N]\}$, constructed recursively by

$$\xi_i^{(0)}(a) = \left\| \exp(r_i^{(0)}/\tau) \right\|_1 \cdot \pi_i^{(0)}(a), \tag{10a}$$

$$\xi_i^{(t+1)}(a) = \xi_i^{(t)}(a)^{1-\eta\tau} \exp(\eta r_i^{(t)}(a)). \tag{10b}$$

Compared with the independent NPG update rule (5), it is clear that $\xi_i^{(t)} \propto \pi_i^{(t)}$ up to normalization. In addition, we have

$$\log \xi_i^{(t+1)} - r_i^{(t+1)}/\tau = (1 - \eta\tau) \log \xi_i^{(t)} + \eta r_i^{(t)} - r_i^{(t+1)}/\tau$$
$$= (1 - \eta\tau) \left(\log \xi_i^{(t)} - r_i^{(t)}/\tau \right) + \left(r_i^{(t)} - r_i^{(t+1)} \right)/\tau,$$

which implies

$$\left\| \log \xi_i^{(t+1)} - r_i^{(t+1)}/\tau \right\|_\infty \leq (1 - \eta\tau) \left\| \log \xi_i^{(t)} - r_i^{(t)}/\tau \right\|_\infty + \left\| r_i^{(t)} - r_i^{(t+1)} \right\|_\infty/\tau$$

$$\leq (1 - \eta\tau)^{t+1} \left\| \log \xi_i^{(0)} - r_i^{(0)}/\tau \right\|_\infty + \tau^{-1} \sum_{s=0}^{t} (1 - \eta\tau)^{t-s} \left\| r_i^{(s)} - r_i^{(s+1)} \right\|_\infty$$

$$\leq (1 - \eta\tau)^{t+1} \left\| \log \xi_i^{(0)} - r_i^{(0)}/\tau \right\|_\infty + \tau^{-1} \sum_{s=0}^{t} (1 - \eta\tau)^{t-s} \sqrt{J\left(\pi^{(s+1)}, \pi^{(s)} \right)}, \quad (11)$$

where the last line follows by applying Lemma 4.2 to the last term by setting $\pi = \pi^{(t)}$ and $\pi' = \pi^{(t+1)}$.

4.3 Step 3: bound the gap

Note that by the definition of the best-response policy in (6), the term of interest in $\mathtt{QRE\text{-}gap}_\tau^{(t)}$ can be controlled as

$$
\begin{aligned}
u_{i,\tau}(\pi_i^{\star(t+1)}, \pi_{-i}^{(t+1)}) - u_{i,\tau}(\pi_i^{(t+1)}, \pi_{-i}^{(t+1)}) &= \big\langle \pi_i^{\star(t+1)} - \pi_i^{(t+1)},\, r_i^{(t+1)} \big\rangle + \tau\mathcal{H}(\pi_i^{\star(t+1)}) - \tau\mathcal{H}(\pi_i^{(t+1)}) \\
&= \tau\mathsf{KL}\big(\pi_i^{(t+1)} \,\|\, \pi_i^{\star(t+1)}\big) \le \tau \left\| \log\pi_i^{(t+1)} - \log\pi_i^{\star(t+1)} \right\|_\infty \\
&\le 2\tau \left\| \log\xi_i^{(t+1)} - r_i^{(t+1)}/\tau \right\|_\infty,
\end{aligned}
$$

where the first line follows from the definition (2), the second step results from a direct consequence of (7):

$$
\big\langle \pi_i^{\star(t+1)} - \pi_i^{(t+1)},\, r_i^{(t+1)} \big\rangle = \big\langle \pi_i^{\star(t+1)} - \pi_i^{(t+1)},\, \tau\log\pi_i^{\star(t+1)} \big\rangle
$$

with a little algebra, and the last line follows from Lemma 4.1. Taking maximum over $i \in [N]$, in conjunction with (11), we end up with

$$
\mathtt{QRE\text{-}gap}_\tau^{(t+1)} \le 2\tau(1-\eta\tau)^{t+1} \left\| \log\pi^{(0)} - \log\pi^{\star(0)} \right\|_\infty + 2\sum_{s=0}^{t}(1-\eta\tau)^{t-s}\sqrt{J\big(\pi^{(s+1)}, \pi^{(s)}\big)}.
$$

Summing the inequality over $t = 0, \ldots, T-1$ gives

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \mathtt{QRE\text{-}gap}_\tau^{(t+1)} \\
&\le 2\tau\sum_{t=0}^{T-1}(1-\eta\tau)^{t+1} \max_{i\in[N]} \left\| \log\pi_i^{(0)} - \log\pi_i^{\star(0)} \right\|_\infty + 2\sum_{t=0}^{T-1}\sum_{s=0}^{t}(1-\eta\tau)^{t-s}\sqrt{J\big(\pi^{(s+1)}, \pi^{(s)}\big)} \\
&\le \frac{2}{\eta\tau}\left(\tau\left\| \log\pi^{(0)} - \log\pi^{\star(0)} \right\|_\infty + \sum_{s=0}^{T-1}\sqrt{J\big(\pi^{(s+1)}, \pi^{(s)}\big)}\right).
\end{aligned}
$$

The proof is thus completed by noticing

$$
\sum_{s=0}^{T-1}\sqrt{J\big(\pi^{(s+1)}, \pi^{(s)}\big)} \le \sqrt{T\sum_{s=0}^{T-1} J\big(\pi^{(s+1)}, \pi^{(s)}\big)} \le \sqrt{2\eta T(\Phi_\tau^{(T)} - \Phi_\tau^{(0)})}.
$$

Here, the second step results from Pinsker's inequality, and the last line follows from (9).

5 Numerical Experiments

We examine the performance of independent NPG methods — in comparison with policy gradient (PG) methods with direct parametrization — on a potential game with $N = 4$ agents and an action space with $|\mathcal{A}| = 20$ actions. The potential function $\Phi(\boldsymbol{a})$ is independently drawn from a Beta distribution $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ for each $\boldsymbol{a} \in \mathcal{A}^N$. We set the learning rate as $\eta = 1/2(\sqrt{N} + \tau)$ for the independent NPG methods, while PG with direct parametrization adopts $\eta = 1/(2N|\mathcal{A}|)$, which is the maximum possible learning rate prescribed in [ZRL21, LOPP21, DWZJ22]. Figure 1 verifies the monotonic improvement of regularized potential function $\Phi_\tau^{(t)}$ for the NPG method, corroborating our theoretical analysis.

Figure 2 shows the averaged performance of various methods over 10 independent runs, in terms of finding approximate NE for the unregularized potential games, and Figure 3 plots that of finding QRE. While independent NPG with a larger regularization parameter $\tau = 10^{-2}$ converges to QRE faster, its $\mathtt{NE\text{-}gap}$ stalls due to the presence of large entropy regularization. With $\tau = 10^{-3}$, NPG achieves a better trade-off and finds a much smaller $\mathtt{NE\text{-}gap}$.
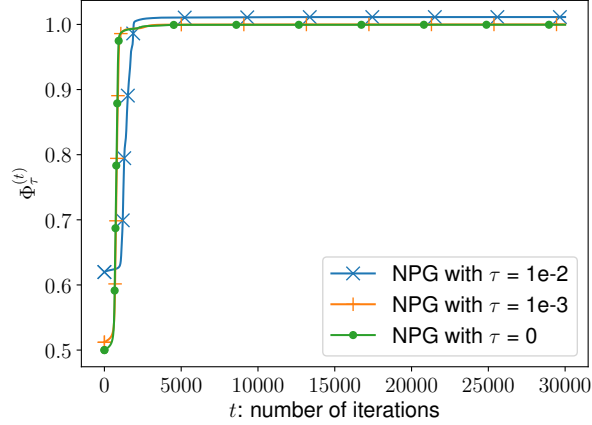
Figure 1: The regularized potential function $\Phi_\tau^{(t)}$ versus the iteration count of NPG with various entropy regularization parameter τ and softmax parameterization.
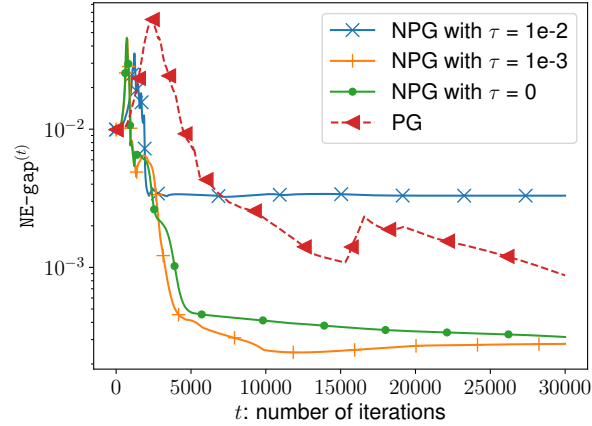

Figure 2: $\texttt{NE-gap}^{(t)}$ versus the iteration count of unregularized PG with direct parameterization and NPG with various entropy regularization parameter τ and softmax parameterization.

6 Conclusions and Discussions

This paper studies independent NPG methods for entropy-regularized potential games and develops a sublinear rate of convergence to quantum response equilibrium, which is independent of the size of the action spaces up to logarithmic factors and grows only sublinearly with respect to the number of agents. In addition, the method achieves the first *dimension-free* convergence rate for the important special case of identical-interest games, where the rate is independent of both the size of the action space and the number of agents. The approach can also be used as a smoothing technique to find Nash equilibria by setting the regularization parameter sufficiently small, without imposing the isolation assumption as often required in prior works. This work leaves open a number of interesting questions:

- Can we tighten the convergence rate in terms of the dependencies on ϵ?

- Can we extend the analysis to establish finite-time global convergence for Markov potential games?

We leave the answers to future work.

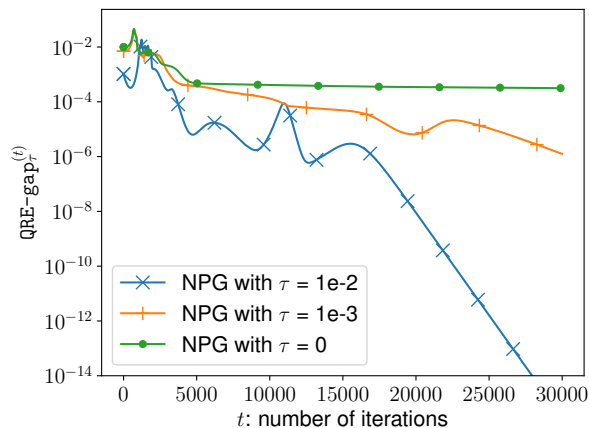Figure 3: $\mathtt{QRE\text{-}gap}_\tau^{(t)}$ versus the iteration count of NPG with various entropy regularization parameter τ and softmax parameterization.

Acknowledgments

References

[ADF+21]   I. Anagnostides, C. Daskalakis, G. Farina, M. Fishelson, N. Golowich, and T. Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. *arXiv preprint arXiv:2111.06008*, 2021.

[AKLM20]   A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

[BR19]   J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[CCC+21]   S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

[CWC21]   S. Cen, Y. Wei, and Y. Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

[Das13]   C. Daskalakis. On the complexity of approximating a Nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

[DFG20]   C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5527–5540, 2020.

[DFG21]   C. Daskalakis, M. Fishelson, and N. Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34, 2021.

[Dri07]   B. K. Driver. Math 280 (probability theory) lecture notes, April 2007. URL: https://mathweb.ucsd.edu/~bdriver/280_06-07/Lecture_Notes/N18_2p.pdf.

[Dur18]   S. Durand. *Analysis of Best Response Dynamics in Potential Games*. PhD thesis, Université Grenoble Alpes, 2018.

[DWZJ22]   D. Ding, C.-Y. Wei, K. Zhang, and M. R. Jovanović. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. *arXiv preprint arXiv:2202.04129*, 2022.

[FGKM18]   M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

[FMOP21]   R. Fox, S. McAleer, W. Overman, and I. Panageas. Independent natural policy gradient always converges in Markov potential games. *arXiv preprint arXiv:2110.10614*, 2021.

[HCM17]   A. Heliou, J. Cohen, and P. Mertikopoulos. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30, 2017.

[HZAL18]   T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

[Jef98]   H. Jeffreys. *The theory of probability*. OUP Oxford, 1998.

[JLWY21]   C. Jin, Q. Liu, Y. Wang, and T. Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent RL. *arXiv preprint arXiv:2110.14555*, 2021.

[Kak01]   S. M. Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[Lan21]   G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

[LFDA16]   S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[LOPP21]   S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

[LWC+21]   G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Softmax policy gradient methods can take exponential time to converge. *arXiv preprint arXiv:2102.11270*, 2021.

[LWT+17]   R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[LZBY20]   Y. Liu, K. Zhang, T. Basar, and W. Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020.

[MB22]   W. Mao and T. Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, pages 1–22, 2022.

[MGY+21]   A. Mirhoseini, A. Goldie, M. Yazgan, J. W. Jiang, E. Songhori, S. Wang, Y.-J. Lee, E. Johnson, O. Pathak, A. Nazi, et al. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021.

[MP95]   R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

[MS96a]   D. Monderer and L. S. Shapley. Fictitious play property for games with identical interests. *Journal of economic theory*, 68(1):258–265, 1996.

[MS96b]   D. Monderer and L. S. Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996.

[MS16]   P. Mertikopoulos and W. H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

[MXD⁺20]   J. Mei, C. Xiao, B. Dai, L. Li, C. Szepesvári, and D. Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33, 2020.

[MXSS20]   J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[Nas51]   J. Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[PPP17]   G. Palaiopanos, I. Panageas, and G. Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. *Advances in Neural Information Processing Systems*, 30, 2017.

[Sel89]   R. Selten. Evolution, learning and economic behavior. 1989.

[SHM⁺16]   D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[SMB21]   Z. Song, S. Mei, and Y. Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

[SV08]   A. Skopalik and B. Vöcking. Inapproximability of pure Nash equilibria. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 355–364, 2008.

[WCYW19]   L. Wang, Q. Cai, Z. Yang, and Z. Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

[WLZL21]   C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. *arXiv preprint arXiv:2102.04540*, 2021.

[Xia22]   L. Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

[ZCH⁺21]   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. *arXiv preprint arXiv:2105.11066*, 2021.

[ZMD⁺22]   R. Zhang, J. Mei, B. Dai, D. Schuurmans, and N. Li. On the effect of log-barrier regularization in decentralized softmax gradient play in multiagent systems. *arXiv preprint arXiv:2202.00872*, 2022.

[ZRL21]   R. Zhang, Z. Ren, and N. Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021.

[ZTLD21]   Y. Zhao, Y. Tian, J. D. Lee, and S. S. Du. Provably efficient policy gradient methods for two-player zero-sum Markov games. *arXiv preprint arXiv:2102.08903*, 2021.

A   Proof of Lemma 4.2

Given any $\pi, \pi' \in \Delta(\mathcal{A})^N$, we have

$$
\begin{aligned}
\left| r_i^\pi(a) - r_i^{\pi'}(a) \right| &= \left| \mathbb{E}_{a_{-i} \sim \pi_{-i}} \left[u_i(a, \, a_{-i}) \right] - \mathbb{E}_{a_{-i} \sim \pi'_{-i}} \left[u_i(a, \, a_{-i}) \right] \right| \\
&\overset{(i)}{\leq} 2 \left\| u_i \right\|_\infty d_{TV} \left(\pi_{-i}, \, \pi'_{-i} \right) \\
&\leq \sqrt{2 d_{TV} \left(\pi_{-i}, \, \pi'_{-i} \right)^2 + 2 d_{TV} \left(\pi'_{-i}, \, \pi_{-i} \right)^2} \\
&\overset{(ii)}{\leq} \sqrt{\mathsf{KL} \left(\pi_{-i} \, \| \, \pi'_{-i} \right) + \mathsf{KL} \left(\pi'_{-i} \, \| \, \pi_{-i} \right)} \\
&\leq \sqrt{\mathsf{KL} \left(\pi \, \| \, \pi' \right) + \mathsf{KL} \left(\pi' \, \| \, \pi \right)} = \sqrt{J(\pi, \pi')},
\end{aligned}
\tag{12}
$$

where $d_{TV}(\cdot, \, \cdot)$ refers to total variation distance. Here, (i) follows from applying $\left| \int_\Omega h \mathrm{d}\mu - \int_\Omega h \mathrm{d}\nu \right| \leq 2 d_{TV}(\mu, \nu) \left\| h \right\|_\infty$ which holds for any probability measures μ, ν and bounded measurable function $h : \Omega \to \mathbb{R}$ (see e.g., [Dri07, Corollary 13.4]), and (ii) results from Pinsker's inequality.

B   Proof of Lemma 4.3

The proof is composed of two parts, each establishing the following bounds

$$
\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \left(\frac{1}{\eta} - \sqrt{N} - \tau \right) J \left(\pi^{(t+1)}, \pi^{(t)} \right),
\tag{13a}
$$

$$
\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \left(\frac{1}{\eta} - 2 \Phi_{\max} - \tau \right) J \left(\pi^{(t+1)}, \pi^{(t)} \right)
\tag{13b}
$$

respectively. Combining the two bounds then finishes the proof.

B.1   Proof of (13a)

We introduce

$$
\widetilde{\pi}_{-i}^{(t)}(a_{-i}) = \prod_{j < i} \pi_j^{(t)}(a_j) \prod_{k > i} \pi_k^{(t+1)}(a_k) \in \Delta(\mathcal{A})^{N-1}
$$

to denote the mixed strategy profile (except that of agent i) where the agents with index $j < i$ follow $\pi_j^{(t)}$ and the agents with index $k > i$ follow $\pi_k^{(t+1)}$ instead. Let $\widetilde{r}_i^{(t)}$ be the associated marginalized utility function, i.e.,

$$
\begin{aligned}
\widetilde{r}_i^{(t)}(a) &= \mathbb{E}_{a_i = a, a_{-i} \sim \widetilde{\pi}_{-i}^{(t)}} \left[u_i(\boldsymbol{a}) \right] \\
&= \sum_{a_{-i} \in \mathcal{A}^{N-1}} u_i(a, \, a_{-i}) \prod_{j < i} \pi_j^{(t)}(a_j) \prod_{k > i} \pi_k^{(t+1)}(a_k).
\end{aligned}
\tag{14}
$$

It follows from the above definition that we have

$$
\Phi_\tau(\pi_i^{(t)}, \widetilde{\pi}_{-i}^{(t)}) = \Phi_\tau(\pi_1^{(t)}, \cdots, \pi_i^{(t)}, \pi_{i+1}^{(t+1)}, \cdots, \pi_N^{(t+1)}) = \Phi_\tau(\pi_{i+1}^{(t+1)}, \widetilde{\pi}_{-(i+1)}^{(t)})
\tag{15}
$$

for $i \in [N-1]$.

We now decompose $\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)}$ as follows:

$$
\begin{aligned}
\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} &= \Phi_\tau(\pi_1^{(t+1)}, \widetilde{\pi}_{-1}^{(t)}) - \Phi_\tau(\pi_N^{(t)}, \widetilde{\pi}_{-N}^{(t)}) \\
&= \Phi_\tau(\pi_1^{(t+1)}, \widetilde{\pi}_{-1}^{(t)}) - \Phi_\tau(\pi_1^{(t)}, \widetilde{\pi}_{-1}^{(t)}) + \Phi_\tau(\pi_1^{(t)}, \widetilde{\pi}_{-1}^{(t)}) - \Phi_\tau(\pi_N^{(t)}, \widetilde{\pi}_{-N}^{(t)})
\end{aligned}
$$

$$\stackrel{(i)}{=} \left[\Phi_\tau(\pi_1^{(t+1)}, \widetilde{\pi}_{-1}^{(t)}) - \Phi_\tau(\pi_1^{(t)}, \widetilde{\pi}_{-1}^{(t)})\right] + \Phi_\tau(\pi_2^{(t+1)}, \widetilde{\pi}_{-2}^{(t)}) - \Phi_\tau(\pi_N^{(t)}, \widetilde{\pi}_{-N}^{(t)})$$

$$\stackrel{(ii)}{=} \sum_{i=1}^{N} \left[\Phi_\tau(\pi_i^{(t+1)}, \widetilde{\pi}_{-i}^{(t)}) - \Phi_\tau(\pi_i^{(t)}, \widetilde{\pi}_{-i}^{(t)})\right]$$

$$= \sum_{i=1}^{N} \left[u_{i,\tau}(\pi_i^{(t+1)}, \widetilde{\pi}_{-i}^{(t)}) - u_{i,\tau}(\pi_i^{(t)}, \widetilde{\pi}_{-i}^{(t)})\right]$$

$$= \sum_{i=1}^{N} \left[\langle \widetilde{r}_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)}\rangle + \tau\left(\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right)\right],$$

where (i) follows from (15), and (ii) follows from repeating the above process over all agents, and the last line follows from (14). For every $i \in [N]$, we have

$$\langle \widetilde{r}_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)}\rangle + \tau\left(\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right)$$
$$= \langle r_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)}\rangle + \tau\left(\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right) + \langle \widetilde{r}_i^{(t)} - r_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)}\rangle.$$

We control the terms separately.

- For the first two terms, recall that taking logarithm on both sides of (5) gives

$$\eta r_i^{(t)} = \log \pi_i^{(t+1)} - (1 - \eta\tau)\log \pi_i^{(t)} + c\mathbf{1}$$

for some constant c. It follows that

$$\langle r_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)}\rangle + \tau\left(\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right) = \frac{1}{\eta}\langle \log \pi_i^{(t+1)} - \log \pi_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)}\rangle$$
$$+ \tau\left(\langle \log \pi_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)}\rangle + \mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right)$$
$$= \left(\frac{1}{\eta} - \tau\right)\mathsf{KL}\left(\pi_i^{(t+1)} \,\|\, \pi_i^{(t)}\right) + \frac{1}{\eta}\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \pi_i^{(t+1)}\right).$$
$$(16)$$

- For the third term, according to (12), we have

$$\left|\widetilde{r}_i^{(t)}(a) - r_i^{(t)}(a)\right| \leq 2d_{TV}(\widetilde{\pi}_{-i}^{(t)}, \pi_{-i}^{(t)}).$$

Hence,

$$\sum_{i=1}^{N} \left|\langle \widetilde{r}_i^{(t)} - r_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)}\rangle\right| \leq 2\sum_{i=1}^{N} d_{TV}(\widetilde{\pi}_{-i}^{(t)}, \pi_{-i}^{(t)})\left\|\pi_i^{(t+1)} - \pi_i^{(t)}\right\|_1$$

$$\stackrel{(i)}{\leq} \frac{2}{\sqrt{N}}\sum_{i=1}^{N} d_{TV}(\pi_{-i}^{(t)}, \widetilde{\pi}_{-i}^{(t)})^2 + \frac{\sqrt{N}}{2}\sum_{i=1}^{N} \left\|\pi_i^{(t+1)} - \pi_i^{(t)}\right\|_1^2$$

$$\stackrel{(ii)}{\leq} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathsf{KL}\left(\pi_{-i}^{(t)} \,\|\, \widetilde{\pi}_{-i}^{(t)}\right) + \sqrt{N}\sum_{i=1}^{N} \mathsf{KL}\left(\pi_i^{(t+1)} \,\|\, \pi_i^{(t)}\right)$$

$$\leq \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathsf{KL}\left(\pi^{(t)} \,\|\, \pi^{(t+1)}\right) + \sqrt{N}\mathsf{KL}\left(\pi^{(t+1)} \,\|\, \pi^{(t)}\right)$$

$$= \sqrt{N}\left(\mathsf{KL}\left(\pi^{(t+1)} \,\|\, \pi^{(t)}\right) + \mathsf{KL}\left(\pi^{(t)} \,\|\, \pi^{(t+1)}\right)\right) = \sqrt{N}J(\pi^{(t+1)}, \pi^{(t)}),$$

where (i) results from Young's inequality and (ii) is due to Pinsker's inequality.

Combining all pieces together, we have

$$\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \left(\frac{1}{\eta} - \tau\right) \sum_{i=1}^N \left[\mathsf{KL}\big(\pi_i^{(t+1)} \,\|\, \pi_i^{(t)}\big) + \mathsf{KL}\big(\pi_i^{(t)} \,\|\, \pi_i^{(t+1)}\big)\right] - \sqrt{N} J(\pi^{(t+1)}, \pi^{(t)})$$

$$\geq \left(\frac{1}{\eta} - \sqrt{N} - \tau\right) J(\pi^{(t+1)}, \pi^{(t)}).$$

B.2 Proof of (13b)

Alternatively, we can decompose $\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)}$ as

$$\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)}$$

$$= \Phi^{(t+1)} - \Phi^{(t)} + \tau \sum_{i=1}^N \left[\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right]$$

$$= \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)} + \tau\mathcal{H}(\pi_i^{(t+1)}) - \tau\mathcal{H}(\pi_i^{(t)})\right] + \Phi^{(t+1)} - \Phi^{(t)} - \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)}\right],$$

$$= \sum_{i=1}^N \left[u_i(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - u_i(\pi^{(t)}) + \tau\mathcal{H}(\pi_i^{(t+1)}) - \tau\mathcal{H}(\pi_i^{(t)})\right] + \Phi^{(t+1)} - \Phi^{(t)} - \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)}\right].$$

The first term is lower bounded by $(1/\eta - \tau)J(\pi_i^{(t+1)}, \pi_i^{(t)})$ as shown in (16). For the remaining terms, we have

$$\left|\Phi^{(t+1)} - \Phi^{(t)} - \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)}\right]\right|$$

$$\leq \sum_{\boldsymbol{a} \in \mathcal{A}^N} \Phi(\boldsymbol{a})\pi^{(t)}(\boldsymbol{a}) \left|\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1 - \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1\right)\right|. \tag{17}$$

To continue, we need the following elementary lemma, which will be proved at the end.

Lemma B.1. *For all $x \in (-1, \infty)$, it holds that*

$$0 \leq x - \log(1 + x) \leq x\log(1 + x).$$

Invoking Lemma B.1 to obtain

$$\left|\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1 - \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1\right)\right|$$

$$= \left|\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1 - \log\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1 - \log\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)}\right)\right|$$

$$\leq \left(\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1\right)\log\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} + \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1\right)\log\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)}.$$

Plugging the above inequality into (17) yield

$$\left|\Phi^{(t+1)} - \Phi^{(t)} - \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)}\right]\right|$$

$$\leq \Phi_{\max} \sum_{\boldsymbol{a} \in \mathcal{A}^N} \pi^{(t)}(\boldsymbol{a}) \left[\left(\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1\right)\log\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} + \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1\right)\log\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)}\right]$$

$$= \Phi_{\max} \sum_{\boldsymbol{a} \in \mathcal{A}^N} \left[\left(\pi^{(t+1)}(\boldsymbol{a}) - \pi^{(t)}(\boldsymbol{a}) \right) \log \frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} \right] + \Phi_{\max} \sum_{i=1}^{N} \sum_{a_i \in \mathcal{A}} \left(\pi_i^{(t+1)}(a_i) - \pi_i^{(t)}(a_i) \right) \log \frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)}$$

$$= \Phi_{\max} \left(J(\pi^{(t+1)}, \pi^{(t)}) + \sum_{i=1}^{N} J(\pi_i^{(t+1)}, \pi_i^{(t)}) \right) = 2\Phi_{\max} J(\pi^{(t+1)}, \pi^{(t)}).$$

Combining all pieces together, we have

$$\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \left(\frac{1}{\eta} - 2\Phi_{\max} - \tau \right) \sum_{i=1}^{N} J(\pi^{(t+1)}, \pi^{(t)}).$$

Proof of Lemma B.1. We have $x - \log(1+x) = x\log(1+x) = 0$ and $(x - \log(1+x))' = (x\log(1+x))' = 0$ when $x = 0$. It follows that $x - \log(1+x) \geq 0$ since $\log$ is concave. With straightforward calculation, we get

$$(x\log(1+x))' - (x - \log(1+x))' = \log(1+x) \begin{cases} \geq 0 & x \geq 0 \\ < 0 & -1 < x < 0 \end{cases},$$

which implies $x - \log(1+x) \leq x\log(1+x)$. $\qquad\square$

B.3 Proof of Corollary 3.2

Proof. By noting that $\pi_i^{\star(0)} \propto \exp\left(r_i^{(0)}/\tau \right)$, and with uniform policy initialization $\pi_i^{(0)} \propto 1$, we can conclude

$$\left\| \log \pi_i^{(0)} - \log \pi_i^{\star(0)} \right\|_\infty \leq 2 \left\| \frac{r_i^{(t)}}{\tau} - 0 \right\|_\infty \leq \frac{2}{\tau}.$$

where the first inequality follows from Lemma 4.1, and the second inequality is true since the payoff is bounded by 1. On the other hand, we have

$$\Phi_\tau^{(T)} - \Phi_\tau^{(0)} = \Phi(\pi^{(T)}) - \Phi(\pi^{(0)}) + \tau\mathcal{H}(\pi^{(T)}) - \tau\mathcal{H}(\pi^{(0)})$$
$$\leq \Phi(\pi^{(T)}) - \Phi(\pi^{(0)}) \leq \Phi_{\max},$$

where the first inequality uses the fact that the entropy is maximized for uniform policies, and the second inequality uses $0 \leq \Phi(\pi) \leq \Phi_{\max}$ for any π. Combining the above two bounds with Theorem 3.1, we have

$$\frac{1}{T} \sum_{t=1}^{T} \texttt{QRE-gap}_\tau^{(t)} \leq \frac{4}{\tau\eta T} + \frac{2}{\tau} \sqrt{\frac{2\Phi_{\max}}{\eta T}}.$$

Setting $\eta = \frac{1}{2(\min\{\sqrt{N}, 2\Phi_{\max}\} + \tau)}$ and $T = O\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}}{\tau^2 \epsilon^2} \right)$ thus completes the proof. $\qquad\square$