Learning in Zero-Sum Markov Games: Relaxing Strong Reachability and Mixing Time Assumptions

Reda Ouhamma, Maryam Kamgarpour

SYCAMORE Lab, École Polytechnique Fédérale de Lausanne (EPFL),
1015 Lausanne, Switzerland

Abstract

We address payoff-based decentralized learning in infinite-horizon zero-sum Markov games. In this setting, each player makes decisions based solely on received rewards, without observing the opponent's strategy or actions, nor sharing information. Prior works established polynomial-time convergence to an approximate Nash equilibrium under strong reachability and mixing time assumptions. We propose a convergent algorithm that significantly relaxes these assumptions, requiring only the existence of a single policy with bounded reachability and mixing time. Our key algorithmic novelty is introducing Tsallis entropy regularization to smooth the best-response policy updates. By suitably tuning this regularization, we ensure sufficient exploration, thus bypassing previous stringent assumptions on the MDP. We prove a polynomial-time convergence to an approximate Nash equilibrium by establishing novel properties of the value and policy updates induced by the Tsallis entropy regularizer.

Introduction

Markov games are a class of multi-agent decision-making problems with a rich history dating back to the foundational work of Shapley (1953). In Markov games, a popular solution concept is a Nash equilibrium, a set of strategies such that no player can improve her payoff by unilaterally changing her strategy. Computing a Nash equilibrium for general Markov games is computationally intractable (Daskalakis, Goldberg, and Papadimitriou 2009; Chen, Deng, and Teng 2009). However, in the special case of zero-sum Markov games, where the players have opposing interests, Nash equilibria can be computed efficiently given known dynamics and rewards. Specifically, Shapley (1953) showed that a stationary Markovian Nash equilibrium can be efficiently computed in discounted zero-sum Markov games. This tractability has led to extensive approaches on computing equilibria in zero-sum Markov games, such as variations of policy and value iterations (Shapley 1953; Hoffman and Karp 1966; Pollatschek and Avi-Itzhak 1969; Van Der Wal 1978; Filar and Tolwinski 1991). Recently, learning in zero-sum Markov games has gained increasing attention, where the goal is learning a Nash equilibrium given only observations of rewards and transitions, without knowing the environment (Littman 1994). The two key objectives in this setting are (Bowling and Veloso 2001): rationality, which requires players to converge to their opponent's best response if the opponent's strategy is asymptotically stationary; and convergence, which refers to reaching a Nash equilibrium.

A particularly desirable framework for learning in Markov games is payoff-based decentralized learning, where players learn without coordinating their actions or observing their opponent's rewards, strategy, and actions. This setting is attractive for its scalability and reduced communication requirements. However, (Liu, Wang, and Jin 2022) proved that learning a Nash equilibrium in this framework is intractable without further assumptions. Therefore, recent works have adopted self-play, where both players use the same learning algorithm. Self-play has been instrumental in solving hard games such as GO, StarCraft, and Dota 2 (Silver et al. 2017; Vinyals et al. 2019; Berner et al. 2019), and has led to theoretical advances in zero-sum Markov games. Existing self-play algorithms with provable convergence to a Nash equilibrium fall into two categories: policy gradient methods, which rely on direct strategy optimization (Wei et al. 2021; Daskalakis, Foster, and Golowich 2020; Zhao et al. 2022; Cen, Wei, and Chi 2021); and value-based methods, which rely on value function estimates to guide strategy updates, including Q-learning-type algorithms (Sayin et al. 2021) and value iteration methods (Chen et al. 2023).

In this paper, we study payoff-based decentralized learning in infinite-horizon discounted zero-sum Markov games. Our goal is to establish polynomial-time convergence guarantees without relying on restrictive assumptions related to reachability and mixing times. In the following, we examine these assumptions, highlighting their roles in prior work and the challenges involved in relaxing them.

Reachability Assumption. A popular assumption on Markov decision processes (MDPs) is strong reachability (Wei, Hong, and Lu 2017; Wei et al. 2021; Chen, Ma, and Zhou 2021; Cai et al. 2024). Namely, there exists a positive L such that *for any strategies* used by players, the expected time to visit any state from any state is bounded by L. Under this assumption, many algorithms demonstrated either average-iterate convergence to an ϵ-approximate Nash

equilibrium (Wei et al. 2021) or last-iterate convergence (Cai et al. 2024). Past efforts that relax strong reachability have weak convergence guarantees. For example, Chen et al. (2023); Sayin et al. (2021) demonstrate convergence up to a bias due to entropy regularization. Specifically, (Sayin et al. 2021) provides an algorithm with an asymptotic best-iterate convergence to a biased Nash equilibrium. The previous result was extended in Chen et al. (2023) to a polynomial-time convergence to a biased Nash equilibrium.

Note that reachability assumptions are related to the difficulty of exploring MDPs. Assuming an episodic setting (Daskalakis, Foster, and Golowich 2020) or access to a generative model (Zhao et al. 2022) bypasses reachability challenges. In the former case, in every new episode, the state is reset and one can avoid being stuck in a state from which reachability is challenged. In the latter, one can query any state and hence, exploration is not an issue.

Mixing-Time Assumption. This common assumption requires a uniform upper bound on the speed of convergence of Markov chains induced by any strategy to their stationary distributions (Chen, Ma, and Zhou 2021). The necessity of this assumption comes from the need to estimate value functions from time-inhomogeneous payoffs, which are due to players changing their strategies with time. This challenge is also relevant for single agent actor-critic algorithms (Xu, Gao, and Gu 2020; Xu, Wang, and Liang 2020; Kumar, Koppel, and Ribeiro 2023; Olshevsky and Gharesifard 2023). For zero-sum Markov games, the algorithm proposed in Chen et al. (2023) also relaxes the above mixing-time assumption, though it only shows convergence to a biased Nash equilibrium as explained above.

The requirement on bounding mixing times arises due to the payoff-based setting as the players need to estimate value functions, and because mixing times bound the speed at which the value functions converge. Note that this requirement would not be needed if players could coordinate to fix their policies when querying the MDP (Wei et al. 2021; Chen, Ma, and Zhou 2021).

Building on this discussion of the objectives and past works in the zero-sum Markov games, we formulate the aim of this paper through the following question:

In zero-sum Markov games, can we learn an approximate Nash equilibrium efficiently without assumptions of strong reachability and uniform mixing times?

Contributions. We address payoff-based decentralized zero-sum Markov games, focusing on the fundamental problem of learning an approximate Nash equilibrium, weakening the above two assumptions. In particular, we only assume the existence of a single strategy that induces an irreducible Markov chain with a finite mixing time, significantly weakening strong reachability and uniform mixing-time assumptions. Our contributions include

- A decentralized, convergent, and rational algorithm;
- A polynomial sample complexity for learning an approximate Nash equilibrium.

Our key contribution is using Tsallis-entropy regularization for zero-sum Markov games. Originally derived from statistical physics, Tsallis entropy generalizes Shannon entropy (Tsallis 1988) and has gained increasing attention in the online learning community for its simultaneous effectiveness in stochastic and adversarial settings (Zimmert and Seldin 2021). This work introduces Tsallis entropy within the smoothed best-response algorithm of (Chen et al. 2023), demonstrating distinct advantages in Markov games due to its improved exploration properties. Specifically, we show:

- Lower bounds on the strategies of our algorithm, ensuring sufficient exploration (see Lemma 1) and polynomial upper bounds on the mixing times;
- The smoothness and strong convexity of a Tsallis-entropy regularized operator, enabling the convergence of our strategies (see Lemma 3).

Organization. In Section 2, we introduce the setting, establish the notation, and discuss the limiting assumptions of past work. Section 3 presents our algorithm and highlights our contributions, including a theorem with the convergence rate to an approximate Nash equilibrium, and corollaries of polynomial sample complexity and rationality. Finally, Section 4 provides a proof sketch outlining the high-level arguments and emphasizing our key technical contributions.

Preliminaries

Notations. We denote the sets of real and natural numbers by $\mathbb{R}$ and $\mathbb{N}$, respectively. Player i's ($i \in \{1,2\}$) opponent is denoted by $-i$. The probability simplex over a finite space $\mathcal{X}$ is denoted by $\Delta^{\mathcal{X}}$.

Problem Setting

We consider infinite-horizon two-player zero-sum Markov games defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, P, \mathcal{R}^1, \mathcal{R}^2, \gamma)$. $\mathcal{S}$ is a finite state space and $\mathcal{A}^i$ is the finite action space of player $i \in \{1,2\}$. We define $A_{\max} = \max(|\mathcal{A}^1|, |\mathcal{A}^2|)$. The transition kernel $P : \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \mapsto \Delta_{\mathcal{S}}$ specifies the probability $P(s'|s, a^1, a^2)$ of transition from s to s' with actions a^1 and a^2. The reward function is $\mathcal{R} : \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \mapsto \mathbb{R}$ for player 1 and $-\mathcal{R}$ for player 2, with $\max_{s,a^1,a^2} |\mathcal{R}(s, a^1, a^2)| \leq 1$. The discount factor γ satisfies $0 < \gamma < 1$. We assume a fixed initial state s_0, without loss of generality (Fiechter 1994).

Payoff-Based Information. We consider a payoff-based decentralized information setting, where players follow the same decision-making algorithm (self-play). In this setup, each player makes decisions based only on her obtained rewards. Moreover, players do not know the other player's strategy or actions and do not share information. At time k, the players observe the state s_k, choose actions $(a_k^i)_{i=1,2}$, and the state transitions to s_{k+1}. Subsequently, the players observe their respective rewards $\mathcal{R}(s_k, a_k^i, a_k^{-i})$ and $-\mathcal{R}(s_k, a_k^i, a_k^{-i})$, and update their strategies independently.

A stationary strategy for player $i \in \{1,2\}$ is a mapping π^i from S to $\Delta^{\mathcal{A}^i}$. We denote by π the joint strategy (π^1, π^2).

We now define the q-functions:

$$q_\pi^1\left(s,a^1\right)=\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}\left(s_t,a_t^1,a_t^2\right)\mid s_0=s,a_0^1=a^1\right],$$

$$q_\pi^2\left(s,a^2\right)=-\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}\left(s_t,a_t^1,a_t^2\right)\mid s_0=s,a_0^2=a^2\right],$$

where the expectation is over the randomness of the strategy π and of the transitions. Finally, the value function of player i, $i=1,2$, is $v_\pi^i(s):=\mathbb{E}_{a^i\sim\pi^i(\cdot|s)}\left[q_\pi^i\left(s,a^i\right)\right]$.

Definition 1. *(Nash equilibrium) The strategies (π_{NE}^1,π_{NE}^2) are a Nash equilibrium (NE) if for $i\in\{1,2\}$, $s\in\mathcal{S}$:*

$$v^i_{\pi^i_{NE},\pi^{-i}_{NE}}(s)\geq v^i_{\pi^i,\pi^{-i}_{NE}}(s),\quad\forall\pi^i\in(\Delta^{\mathcal{A}_i})^{\mathcal{S}},$$

meaning that no agent can improve their value by unilaterally changing their strategy.

Definition 2. *(Approximate Nash equilibrium) Given a strategy (π^i,π^{-i}), we define the Nash Gap:*

$$\mathrm{NG}(\pi^i,\pi^{-i})=\max_{s\in\mathcal{S}}\left\{\sum_{i=1,2}\left(\max_{\hat{\pi}^i\in\Delta^{\mathcal{A}^i}}v^i_{\hat{\pi}^i,\pi^{-i}}(s)-v^i_{(\pi^i,\pi^{-i})}(s)\right)\right\}$$

For an $\epsilon>0$, the strategies (π^i,π^{-i}) are an ϵ-approximate Nash equilibrium if $\mathrm{NG}(\pi^i,\pi^{-i})\leq\epsilon$.

The Nash Gap equals zero if and only if the strategies constitute a Nash equilibrium. Our objective is to learn an approximate Nash equilibrium in polynomial time without assumptions of strong reachability or uniform mixing times. There exist other notions of convergence, such as average-iterate convergence and no-regret. Average-iterate entails a bounded average Nash gap over strategies played throughout the iterations of the algorithm. A no-regret result implies that each player's regret is sublinear given the opponent's played strategies over the iterates of the algorithm. In normal-form games, no-regret implies average-iterate convergence to a Nash equilibrium (Freund and Schapire 1999) but not last-iterate convergence (Muthukumar, Phade, and Sahai 2020). Thus, the notion of convergence we are after is stronger than no-regret and average-iterate convergence.

Limiting Assumptions In State-Of-The-Art

Strong Reachability. We begin with the common strong reachability assumption (Auer and Ortner 2006; Chen, Ma, and Zhou 2021), also known as the irreducible game assumption in (Cai et al. 2024).

Definition 3. *(Strong reachability) The MDP satisfies strong reachability if there exists a finite constant $L>0$ such that:*

$$\max_{s,s'\in\mathcal{S}}\max_{\pi\in(\Delta^{\mathcal{A}_i})^{\mathcal{S}}\times(\Delta^{\mathcal{A}_{-i}})^{\mathcal{S}}}T^\pi_{s\mapsto s'}\leq L,$$

where $T^\pi_{s\mapsto s'}$ is the expected time to reach state s' from state s when players follow strategy π.

In particular, this means that strong reachability implies that the Markov chain induced by any strategy is irreducible: any two states are reachable from each other by a finite sequence of transitions with positive probability. Hence, this is an assumption on how easy it is to explore an MDP. It was shown in Durrett (2019, Theorem 5.5.11) that an irreducible Markov chain induced by a strategy π with stationary distribution μ_π satisfies

$$T^\pi_{s\mapsto s}=1/\mu_\pi(s).\tag{1}$$

Therefore, if the Markov chain induced by any strategy π is irreducible, then strong reachability is equivalent to the stationary distribution being uniformly lower bounded $\mu_\pi(s)\geq 1/L$ over states and strategies. The assumption of a uniform lower bound on the stationary distribution over strategies is prevalent in reinforcement learning (Agarwal et al. 2021; Mei et al. 2020; Zhang, Ren, and Li 2022).

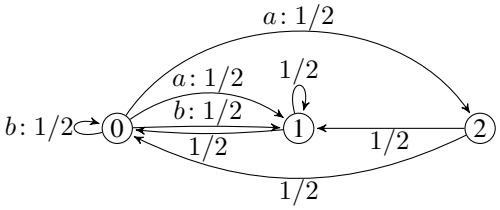

Figure 1: MDP with three states: transitions from states 1 and 2 are action independent, two possible actions a and b in state 0. Arrows indicate transitions and their probabilities.

To see the limitation of strong reachability, consider the MDP in Figure 1, and consider the strategy π parameterized by $\xi\in[0,1]$ defined as:

$$\pi(0,a)=\xi\text{ and }\quad\pi(0,b)=1-\xi.\tag{2}$$

Then, the corresponding stationary distribution μ_π is:

$$\mu_\pi=\left(\frac{1}{2+\xi},\frac{1}{2},\frac{\xi}{4+2\xi}\right),\tag{3}$$

which implies $\lim_{\xi\to 0}\mu_\pi(2)=0$. This invalidates Assumption 3 because there couldn't exist a positive L such that $\min_s\inf_\pi\mu_\pi(s)\geq 1/L$, see Equation (1).

In single-agent reinforcement learning, Auer, Jaksch, and Ortner (2008) proposed a weaker alternative assumption. Namely, Auer, Jaksch, and Ortner (2008) proved minimax optimal bounds requiring only that, for each pair of states, there exists a strategy under which the expected time to travel between them is bounded: $\max_{s,s'}\min_\pi T^\pi_{s\mapsto s'}\leq L$.

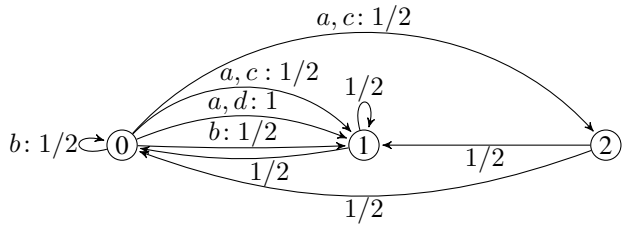

Figure 2: Two-player MDP with three states: augments the MDP of Figure 1, the second player has two actions c and d in state 0. Arrows indicate transitions and their probabilities.

In multi-agent settings, relaxing strong reachability is more challenging because players can block each other from

reaching certain states. For example, in the MDP of Figure 2, the second player can prevent player 1 from reaching state 2 by consistently choosing action d. This highlights that a game setting exacerbates reachability challenges compared to a single-agent setting. To our knowledge, all previous works proving convergence to a Nash equilibrium in zero-sum Markov games have required strong reachability.

Uniform Mixing Times. This assumption asserts the existence of a uniform upper-bound, with respect to the strategies, on the mixing times of the Markov chain induced by any strategy (Olshevsky and Gharesifard 2023; Chen and Zhao 2024; Bhandari, Russo, and Singal 2018; Wu et al. 2020). To state this assumption, we introduce two necessary notations. For a strategy $\pi = (\pi_1, \pi_2)$ and a transition matrix P, we define P_π as the transition kernel induced by strategy π, and P_π^k as the k-step transition kernel for $k \in \mathbb{N}$.

Definition 4. *The ϵ-mixing time of strategy π, with unique stationary distribution μ_π, is defined as:* $t_{\pi,\epsilon} = \min\{k \geq 0 : \max_{s\in\mathcal{S}} \left\|P_\pi^k(\cdot|s) - \mu_\pi(\cdot)\right\|_{\mathrm{TV}} \leq \epsilon\}$, *where $\|.\|_{\mathrm{TV}}$ is the total variation distance.*

Definition 5. *(Assumption of uniform mixing-time) The MDP satisfies the uniform mixing-time assumption if for any $\epsilon > 0$, there exists a finite mixing-time $t_{mix}(\epsilon)$ such that*

$$\forall \pi \in (\Delta^{\mathcal{A}_i})^{\mathcal{S}} \times (\Delta^{\mathcal{A}_{-i}})^{\mathcal{S}}, \quad t_{\pi,\epsilon} \leq t_{mix}(\epsilon).$$

Assuming a finite bound on the supremum of the mixing times over strategies, as in Definition 5, is very restrictive. To illustrate, we show that for the MDP in Figure 1 with the strategies defined in Equation (2), it holds that the mixing time grows to infinity when ξ tends to one (see Appendix E of (Ouhamma and Kamgarpour 2023)). This invalidates the assumption in Definition 5 because the mixing times are not uniformly bounded.

In single-agent reinforcement learning, the assumption of uniform mixing time can be relaxed for algorithms that perform many value function updates per strategy update, see (Kumar, Koppel, and Ribeiro 2023). Intuitively, this assumption characterizes how fast value function estimates achieve their steady-state value. While this approach is acceptable in a single-agent setting, applying it in a multi-agent context would require significant coordination between the players to synchronize their updates.

Algorithm and Sample Complexity

Algorithm

We present our algorithm, Tsallis smoothed Best-Response Dynamics with Value Iteration, `TBRVI` . It builds on the algorithm of Chen et al. (2023), which combines principles of value iteration and best-response dynamics. Our primary contribution lies in introducing Tsallis entropy regularization for strategy updates, replacing the softmax smoothing (Shannon entropy) used in previous work.

Algorithm Statement. The pseudo-code of `TBRVI` is presented in Algorithm 1. It takes as input the number of episodes T, the length of an episode K, and a regularization parameter η.

Algorithm 1: Tsallis-smoothed Best-Response Dynamics with Value Iteration

1: **Input:** Integers K and T, real number $\eta > 0$, matrices $v_0^i = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}|}$, $q_{t,0}^i = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|}$ for all t, strategies $\pi_{t,0}^i(a^i|s) = 1/|\mathcal{A}^i|$ for all (s, a^i) and t.
2: **for** $t = 0, 1, \cdots, T$ **do**
3: **for** $k = 0, 1, \cdots, K-1$ **do**
4: Update strategies: $\forall s \in \mathcal{S}, i \in \{1, 2\} : \pi_{t,k+1}^i(s) = \pi_{t,k}^i(s) + \beta_k(\mathrm{Ts}(q_{t,k}^i(s, .)) - \pi_{t,k}^i(s))$
5: Sample actions: $a_k^i \sim \pi_{t,k+1}^i(\cdot|s_k)$
6: Observe: $s_{k+1} \sim P(\cdot \mid s_k, a_k^i, a_k^{-i})$
7: Update values, for all (s, a^i):

$$q_{t,k+1}^i(s, a^i) = q_{t,k}^i(s, a^i) + \alpha_k \Big(\mathcal{R}^i(s_k, a_k^i, a_k^{-i}) + \gamma v_t^i(s_{k+1}) - q_{t,k}^i(s_k, a_k^i) \Big) \mathbb{1}_{\{s=s_k, a^i=a_k^i\}}$$

8: **end for**
9: $v_{t+1}^i(s) = \pi_{t,K}^i(s)^\top q_{t,K}^i(s, .)$ for all $s \in \mathcal{S}$ and set $s_0 = s_K$
10: **end for**
11: **Output:** $\pi_{T,K}^i$

In Algorithm 1, the value functions and the q-functions are initialized at zero. In line 4, the strategies are updated according to a smoothed best response. The smoothing is based on Tsallis entropy regularization with $\alpha = 1/2$, see Tsallis (1988), leading to the following strategy update:

$$\mathrm{Ts}(q_{t,k}^i(s)) = \underset{w \in \Delta^{|\mathcal{A}^i|}}{\arg\max} \left\langle w, q_{t,k}^i(s, .) \right\rangle + \frac{4}{\eta} \sum_j \sqrt{w_j}\,, \quad (4)$$

where $\eta > 0$ is a learning rate. Note that without regularization, the update becomes a maximization of $\left\langle w, q_{t,k}^i(s, .) \right\rangle$ which is equal to $\mathbb{E}_{a^i \sim w}[q_{t,k}^i(s, .)]$, similar to the best-response dynamics (Hofbauer and Sorin 2006). Finally, Zimmert and Seldin (2021) provided a closed form for the Tsallis update: $\mathrm{Ts}(q_{t,k}^i(s)) = (4\eta^{-2}(q_{t,k}^i(s, a) - x)^{-2})_{a\in\mathcal{A}}$, where the constant $x \in \mathbb{R}$ is defined implicitly through the normalization constraint $\sum_{a\in\mathcal{A}} 4\left(\eta\left(q_{t,k}^i(s, a) - x\right)\right)^{-2} = 1$.

Next, in line 5, the action of player i is sampled from strategy $\pi_{t,k+1}^i(.|s_k)$. Subsequently, in line 6, the MDP transitions to state s_{k+1} according to the transition dynamics. Line 7 depicts the q-functions' updates. Finally, in line 9, the value functions are updated similarly to value-iteration. Note that, unlike episodic settings, the state is not reset at the end of an episode.

The inner loop of `TBRVI` (lines 4 to 7) must ensure the convergence of strategies and q-functions, and this presents a key challenge compared to past settings and analyses. The convergence of value updates in the outer loop, line 9, typically requires strategies to visit all state actions sufficiently (Sutton and Barto 2018). Past works ensure this requirement by assuming strong reachability and using Shannon en-

tropy, through softmax smoothing. By introducing the Tsallis entropy, we provide better reachability bounds, as we will prove. Intuitively, Tsallis entropy entails more exploration of suboptimal actions (Lemma 1) and leads to faster mixing times (Lemma 6 of (Ouhamma and Kamgarpour 2023)).

Theoretical Statement

Here, we show TBRVI 's polynomial-time convergence to an approximate Nash equilibrium. Rather than strong reachability and uniform mixing-time assumptions, we consider a relaxed assumption.

Assumption 1. *There* exists *a joint strategy $\pi_r = (\pi_r^i, \pi_r^{-i})$ such that P_{π_r} induces a Markov chain which is irreducible and has a finite mixing-time.*

The irreducibility here is strictly weaker than strong reachability (Definition 3) as it only concerns a single strategy instead of any possible strategy. Similarly, the existence of a finite mixing time for the strategy π_r relaxes the assumption of uniformly bounded mixing times (Definition 5). For example, the two-player MDP discussed after Equation (3) satisfies Assumption 1 with any strategy in which the second player chooses action d with positive probability. However, this MDP does not satisfy strong reachability or uniformly bounded mixing times as we previously showed.

We are now ready to present our convergence result. This is based on several key characteristics of the reference strategy π_r of Assumption 1. First, the reachability here is measured by the quantity $\mu_{r,\min} = \min_{s\in\mathcal{S}} \mu_r(s)$, where μ_r is the stationary distribution of strategy π_r. The quantity $\mu_{r,\min}$ is guaranteed to be positive by Assumption 1 and Equation (1). Lastly, our convergence rate depends on the mixing time of π_r, characterized by the spectral gap $\lambda_r = 1 - \lambda_2$, where λ_2 is the second largest eigenvalue of P_{π_r}.

Theorem 1 (Nash Gap bound)**.** *Assume that the players follow Algorithm 1 with $\alpha_k = \frac{\alpha}{k+h}$ and $\beta_k = \frac{\beta}{k+h}$. In addition, define $\ell_\eta = \frac{1}{\left(\sqrt{A_{\max}}+\frac{\eta}{2(1-\gamma)}\right)^2}$, $c_\eta = \mu_{r,\min}\ell_\eta^3$, and choose $\beta > 4$, $\alpha > \frac{1}{\ell_\eta \mu_{r,\min}}$, and $h > \beta$ such that $\frac{\beta}{\alpha} \leq \min\left\{\frac{c_\eta \ell_\eta^3 (1-\gamma)^2}{6272\eta^3|\mathcal{S}|A_{\max}^4}, \frac{c_\eta(1-\gamma)^2}{34|S|A_{\max}\mu_{r,\min}}, \frac{1}{60|S|A_{\max}L_\eta}\right\}$, where L_η is defined in Lemma 4 of (Zhang et al. 2023). Then, under Assumption 1, it holds for all $K \geq k_0$:*

$$\mathbb{E}[NG(\pi_{T,K}^i, \pi_{T,K}^{-i})] \leq \frac{c_1|\mathcal{S}|A_{\max}T\eta}{(1-\gamma)^3}\left(\frac{\gamma+1}{2}\right)^{T-1} + \frac{c_2|\mathcal{S}|A_{\max}\tau_K^2\alpha^{3/2}}{\alpha_{k_0}^{1/2}\beta(1-\gamma)^4}\frac{1}{\sqrt{K}} + \frac{c_3\sqrt{A_{\max}}}{\eta(1-\gamma)^2},$$

where $k_0 = \min\{k \geq 0 \mid k \geq \tau_k\}$, $\tau_K = t_{\ell_\eta,\beta_k}$, and $\{c_j\}_{1\leq j\leq 3}$ are numerical constants.

Observe that $\ell_\eta = \mathcal{O}(\eta^{-2})$, $\tau_K = \mathcal{O}(\log(K)\eta^{12})$, $L_\eta = \mathcal{O}(\eta^{12})$, and $c_\eta = \mathcal{O}(\eta^{-6})$, and thus, they have polynomial dependence on η. This enables us to establish convergence to an ϵ-approximate Nash equilibrium with polynomial sample complexity in $1/\epsilon$ as will be shown below.

Theorem 1 provides an upper bound on the Nash gap along the iterates of our algorithm. The first term in the inequality represents the bias induced by the minimax value iteration (line 9 of Algorithm 1) and the contraction property of the Bellman operator due to the discount factor. The second term captures the combined convergence error and variance of the inner loop. Its scaling as $1/\sqrt{K}$ is consistent with typical bounds from online learning. The third term corresponds to a regularization bias arising from the use of Tsallis entropy smoothing. By optimizing these three terms with respect to η, we establish the convergence of the iterates to an ϵ-approximate Nash equilibrium.

Corollary 1 (Sample Complexity)**.** *Under Assumption 1 and with $\eta = K^{1/80}$, for any $\epsilon > 0$, TBRVI achieves $\mathbb{E}[NG(\pi_{T,K}^i, \pi_{T,K}^{-i})] \leq \epsilon$ for $K = \widetilde{\mathcal{O}}\left(1/\epsilon^{80}\right)$ and $T = \widetilde{\mathcal{O}}(\log(1/\epsilon))$. Thus, TBRVI learns an ϵ-approximate Nash equilibrium with a polynomial sample complexity in $1/\epsilon$.*

To our knowledge, this is the first polynomial sample complexity for an ϵ-approximate Nash equilibrium without assumptions of reachability or uniform mixing times.

We further show that for a player using TBRVI , her iterates converge in polynomial time to the best response of an opponent playing a stationary strategy.

Corollary 2 (Rationality)**.** *Let player $-i$ follow a strategy π^{-i}. Assume there exists a strategy π_i such that the joint strategy (π^i, π^{-i}) induces an irreducible Markov chain with a finite mixing time. Then for any $\epsilon > 0$, if the player i follows TBRVI with $\eta = K^{1/80}$, then*

$$\max_{\hat{\pi}^i} v^i_{(\hat{\pi}^i, \pi^{-i})}(s_0) - v^i_{(\pi^i_{T,K}, \pi^{-i})}(s_0) \leq \epsilon$$

for $K = \widetilde{\mathcal{O}}\left(1/\epsilon^{80}\right)$ and $T = \widetilde{\mathcal{O}}(\log(1/\epsilon))$.

Observe that the assumption in the statement of Corollary 2 is equivalent to Assumption 1 in the case of a fixed opponent. The above corollary establishes the so-called *rationality* of the algorithm as referred to in (Cen, Wei, and Chi 2021; Chen, Ma, and Zhou 2021; Wei et al. 2021).

Comparison with State-of-the-Art. Let us compare with the closest works addressing payoff-based decentralized learning of equilibria in zero-sum Markov games. In Wei et al. (2021), a sample complexity of $\tilde{\mathcal{O}}\left(1/\epsilon^8\right)$ was derived for the weaker notion of average iterate convergence. Furthermore, the above work required strong reachability and proposed an algorithm that updated strategies and q functions on separate timescales, necessitating a higher level of coordination between players. Specifically, their method involved players iteratively coordinating to fix their strategies while collecting batches of samples to estimate the value functions. For last-iterate convergence, Chen, Ma, and Zhou (2021) derived a $\widetilde{\mathcal{O}}(1/\epsilon^{5.5})$ sample complexity assuming strong reachability and uniformly bounded mixing times. Furthermore, their algorithm required communicating the entropy of the player's strategy to the opponent. Recently, Cai et al. (2024) proved a last-iterate guarantee with a rate of $\tilde{\mathcal{O}}\left(L^{1/\xi}/\epsilon^{9+\xi}\right)$ for any $\xi > 0$ under the assumption of strong reachability. Relaxing this assumption, Cai et al. (2024) showed a so-called sample path convergence, implying that "for all states that players visit often enough, players

learn an approximate Nash strategy". As proven in Cai et al. (2024), this convergence notion is weaker than its average-iterate and last-iterate counterparts. Finally, past efforts relaxing strong reachability include (Sayin et al. 2021; Chen et al. 2023). Specifically, Sayin et al. (2021) proved asymptotic convergence to a biased Nash equilibrium under a condition similar to Assumption 1, and Chen et al. (2023) extended the result to a polynomial-time convergence to a biased Nash equilibrium.

Proof Sketch

Our objective in this section is to provide insight into the proof, highlighting the role of Tsallis entropy in overcoming assumptions of strong reachability and uniformly bounded mixing times. We attribute the prevalence of restrictive assumptions to the popular use of softmax smoothing in the literature. Specifically, softmax smoothing induces strategies that converge too rapidly. Thus, the probability of selecting suboptimal actions decays exponentially fast, hindering the convergence of value functions. Moreover, this decay influences the mixing time bounds, which grow inversely proportional to the level of exploration of the strategies (Chen et al. 2023, Lemma 4).

Key Properties of Tsallis Entropy

Tsallis entropy, $\mathcal{H}_\alpha(\pi) = \frac{1}{1-\alpha}(1-\sum_i \pi_i^\alpha)$, $\alpha \in [0,1]$, generalizes the Shannon entropy and the log-barrier potential as special cases for $\alpha \to 1$ and $\alpha \to 0$, respectively (Agarwal et al. 2017; Abernethy, Lee, and Tewari 2015). The strategies induced by Tsallis entropy have a closed-form expression, see Equation (4). This expression allows us to establish the following crucial lower bound on the strategies deployed by Algorithm 1.

Lemma 1 (Policy lower bound)**.** *For all $t, k \geq 0$, $s \in S, a^i \in \mathcal{A}$, $i \in \{1,2\}$ the following lower bound on the strategies $\pi^i_{t,k}$ of* `TBRVI` *holds: $\pi^i_{t,k}(a^i|s) \geq \ell_\eta$, where $\ell_\eta = 1/\left(\sqrt{A} + \frac{\eta}{2(1-\gamma)}\right)^2$.*

The proof of this lemma is provided in the full paper, see (Ouhamma and Kamgarpour 2023). The above ensures that strategies are lower bounded by a function of the regularization coefficient η. This bound is utilized in the proof of Theorem 1 in two ways. First, we use it to prove a lower bound on the stationary distribution of the strategies of `TBRVI` , see Lemma 2 below; Second, we can bound the mixing time of the strategies, see Lemma 6 of (Ouhamma and Kamgarpour 2023).

Lemma 2 (Stationary distribution lower bound)**.** *Define the strategy class $\Pi_\delta = \{\pi = (\pi^i, \pi^{-i}) \mid \min_{s,a^i} \pi^i(a^i|s) > \delta_i,\ \text{for } i = 1,2\}$, where $\delta_i, \delta_{-i} \in (0,1)$. For any strategy $\pi \in \Pi_\delta$ with stationary distribution μ_π, the following lower bound on μ_π holds: $\mu_\pi(s) \geq \mu_{r,\min}\delta_1\delta_2$.*

The proof of this lemma is provided in Appendix B.3 of (Ouhamma and Kamgarpour 2023). Intuitively, it follows from the observation that a sufficiently exploratory strategy coincides with π_r with some probability. This exploratory behavior is guaranteed by the use of Tsallis entropy regularization. The argument then involves algebraic manipulations and relating the stationary distribution of any exploratory strategy to μ_r. Notably, Lemma 2 also holds for Shannon entropy-regularized strategies; however, the resulting lower bound would be exponentially small, making it impractical for the final analysis.

Using the above-established properties, we proceed to bound the Nash gap through drift inequalities defined below.

Drift Inequalities

The Nash gap can be decomposed as below (Chen et al. 2023, Equation 7).

$$\mathrm{NG}(\pi^i_{T,K}, \pi^{-i}_{T,K}) \leq C_0\Bigg(\mathcal{D}_\pi(T,K) + 2\underbrace{\|v^i_T + v^{-i}_T\|_\infty}_{\mathcal{D}} + \sum_{i=1,2}\underbrace{\|v^i_T - v^i_*\|_\infty}_{\mathcal{D}_i} + \underbrace{\frac{8\sqrt{A_{\max}}}{\eta}}_{\text{bias}}\Bigg), \quad (5)$$

where C_0 is a constant, $\mathcal{D}_\pi(\cdot)$ is a term we refer to as strategy drift below, and v^i_* is the unique fixed-point of a Bellman operator (see Appendix A of (Ouhamma and Kamgarpour 2023)).

Our proof builds on that of Chen et al. (2023), establishing drift inequalities for each term on the right-hand side. Drift inequalities are inequalities showing a negative drift of the iterate (similar to Lyapunov inequalities) with additional terms arising from couplings with other iterates. Let us now highlight how we use Tsallis entropy properties established in Section to establish drift inequalities for $\mathcal{D}_\pi$, $\mathcal{D}$, and $\mathcal{D}_i$, $i = 1,2$.

Policy Drift $\mathcal{D}_\pi$. The first term $\mathcal{D}_\pi$, in the Nash gap decomposition is the sum of the so-called Lyapunov functions (Hofbauer and Hopkins 2005), as defined below.

$$\mathcal{D}_\pi = \sum_s V_{v,s}\left(\pi^i, \pi^{-i}\right),$$

$$V_{v,s}\left(\pi^i, \pi^{-i}\right) = \sum_{i=1,2}\max_{\hat{\pi}^i \in \Delta^{|\mathcal{A}^i|}}\Big\{\left(\hat{\pi}^i - \pi^i\right)^\top \mathcal{T}^i(v^i)(s)\pi^{-i} + \frac{1}{\eta}\left(\mathcal{H}\left(\hat{\pi}^i\right) - \mathcal{H}\left(\pi^i\right)\right)\Big\}, \quad (6)$$

where $\mathcal{T}^i(v)(s,a^i,a^{-i}) = \mathcal{R}^i(s,a^i,a^{-i}) + \gamma\mathbb{E}\left[v(s_1)|s,a^i,a^{-i}\right]$. The function $V_{v,s}$ serves as a regularized Nash gap for the matrix game with payoffs $\mathcal{T}^i(v)(s,\cdot,\cdot)$. It is an adaptation of the Lyapunov function provided in Hofbauer and Hopkins (2005) for best response dynamics in matrix games. In particular, it is adjusted to accommodate the Markov games setting by using $\mathcal{T}^i(v)$ as the payoff matrices. Furthermore, rather than the previously adopted Shannon entropy, the regularization $\mathcal{H}(\cdot)$ is with Tsallis entropy, aligning with our algorithmic choices.

To analyze $\mathcal{D}_\pi$, we prove that $V_{v,s}$ is strongly convex and smooth (see Lemma 11 of (Ouhamma and Kamgarpour 2023)). A key technical contribution that enabled these properties is this Lipschitzness result.

Lemma 3 (Lipschitzness of Tsallis entropy)**.** *For all $\boldsymbol{R}$ and $\boldsymbol{R}' \in \mathbb{R}^n$, we have:*

$$\|\operatorname{Ts}(\boldsymbol{R}) - \operatorname{Ts}(\boldsymbol{R}')\|_2 \leq 2\sqrt{2}\eta\, n \|\boldsymbol{R} - \boldsymbol{R}'\|_2.$$

The proof is provided in Appendix B.1 of (Ouhamma and Kamgarpour 2023). As a consequence, we show that $V_{v,s}$ is smooth and strongly convex, crucial properties for the negative drift of $\mathcal{D}_\pi$. Moreover, an analogous Lipschitz property for Shannon entropy was derived in Gao and Pavel (2017) and has been instrumental in prior convergence analyses, see Chen et al. (2023, Lemma A.7) and Chen, Ma, and Zhou (2021, Lemma 6). Given this significance, we believe the Lipschitz property of Tsallis entropy may also be of independent interest.

Value Function Drifts $\mathcal{D}, \mathcal{D}_i$, $i = 1, 2$. We refer to $\mathcal{D}$ and $(\mathcal{D}_i)_{i=1,2}$ as value function drifts, as they relate to different terms in the estimation of value function $((v_t^i)_{i=1,2})_{t\geq 0}$. For each of these terms, we prove an inequality that exhibits a negative drift with the addition of a coupling term. The negative drift primarily arises from the contractiveness of the Bellman operator, while the coupling terms reflect the dependence on the value function of player i on that of the other player, and the strategy iterates. The analysis of $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}$ relies on properties of Tsallis entropy. For the convergence of value functions in line 7 of Algorithm 1, it is crucial that the strategies appear stationary. To ensure this, we choose sufficiently large episode lengths so that the Markov chains are close to their stationary distribution, thus ensuring accurate value function estimation. To determine the sufficient episode length, we derive a mixing time upper bound for the Markov chains induced by the policies of Algorithm 1, see Lemma 6 of (Ouhamma and Kamgarpour 2023). This mixing time bound is novel and is derived following classical arguments using the so-called spectral gap, see (Montenegro, Tetali et al. 2006).

Bias Term. The bias term $\left(8\sqrt{A_{\max}}/\eta\right)$, is the last term in our decomposition. The objective is to select a sufficiently large η to control this regularization bias. Observe that this bias is also the last term in the Nash gap bound of Theorem 1, whereas the first two terms therein capture the cumulative effects of the drift terms $\mathcal{D}_\pi, \mathcal{D}, \mathcal{D}_1, \mathcal{D}_2$. Importantly, our bound on the drift functions scales polynomially with η (see Theorem 1), allowing us to optimize the Nash gap bound with respect to η, effectively removing the bias and establishing a polynomial sample complexity in $1/\epsilon$.

In contrast, prior approaches using softmax smoothing suffer from exponentially decaying strategy lower bounds and exponentially increasing mixing-time upper bounds, see Auletta et al. (2013). As a result, softmax-smoothed algorithms cannot balance the bias term with the drift terms while maintaining a polynomial convergence rate in $1/\epsilon$. This underscores the critical advantage of Tsallis entropy, which yields a bias term that can be successfully controlled.

Discussion

Implementability of the Algorithm. In Theorem 1 and Corollary 1, the key quantities λ_r and $\mu_{r,\min}$ are required to select the parameters of `TBRVI` . However, these quantities are typically unknown in practice and must be estimated. In single-agent reinforcement learning, several well-established methods exist for approximating such parameters. For instance, the spectral gap λ_r can be estimated using coupling methods, as demonstrated in Jerrum (2003, Section 4.2). Similarly, practical techniques for estimating the stationary distribution of a Markov chain have been developed by Asmussen and Glynn (2007, Chapter IV).

Learning rate. Theorem 1 recommends choosing the stepsizes to satisfy $\beta/\alpha = \mathcal{O}\left(\eta^{-15}/k\right)$. This means that β can be very small in practice. While this is a necessary aspect of the theoretical analysis, it can be limiting in practice. Small learning rates of this nature are common in analyses of algorithms involving simultaneous value and strategy updates, both in single-agent reinforcement learning (Olshevsky and Gharesifard 2023; Konda and Borkar 1999; Khodadadian et al. 2022) and in Markov games (Chen et al. 2023; Cai and Daskalakis 2011). While it is possible to avoid this issue by collecting many samples of the MDP per strategy update (Wei et al. 2021; Cai et al. 2024), implementing such algorithms in a game setting requires significant coordination, as the value function of player i depends on the strategy π^{-i} of the opponent. Hence, the players need to fix their strategies in a coordinated way.

Conclusion

This work addressed learning an approximate Nash equilibrium in zero-sum Markov games by proposing a payoff-based and decentralized algorithm. The main contribution was relaxing strong reachability and uniform mixing time assumptions made in the prior works. Specifically, we introduced a relaxed reachability requirement in Assumption 1 and proposed the `TBRVI` algorithm, which provably converges to an ϵ-approximate Nash in polynomial time in $1/\epsilon$. The key algorithmic contribution was the use of Tsallis entropy to obtain smooth strategy updates.

The work opens up several promising research directions. First, our sample complexity bounds may be improved using concentration inequality-based analyses, which have provided the best convergence rates so far under the assumptions of strong reachability and uniformly bounded mixing times (Wei et al. 2021; Chen, Ma, and Zhou 2021). Second, given the slow learning rates required in our algorithm, an important direction is to explore sample complexity lower bounds for Markov games under the payoff-based information setting and in the absence of coordination. Finally, to address real-world applications, extending both the approach and analysis to continuous state and action spaces, as well as validating them through real-world experiments, is a crucial next step.

Acknowledgements

Reda Ouhamma and Maryam Kamgarpour are funded by an SNSF grant under NCCR Automation and an SNSF grant under #200020_207984/1

References

Abernethy, J. D.; Lee, C.; and Tewari, A. 2015. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 28.

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2021. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98): 1–76.

Agarwal, A.; Luo, H.; Neyshabur, B.; and Schapire, R. E. 2017. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, 12–38. PMLR.

Asmussen, S.; and Glynn, P. W. 2007. *Stochastic simulation: algorithms and analysis*, volume 57. Springer.

Auer, P.; Jaksch, T.; and Ortner, R. 2008. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21.

Auer, P.; and Ortner, R. 2006. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19.

Auletta, V.; Ferraioli, D.; Pasquale, F.; and Persiano, G. 2013. Mixing time and stationary expected social welfare of logit dynamics. *Theory of Computing Systems*, 53: 3–40.

Berner, C.; Brockman, G.; Chan, B.; Cheung, V.; Debiak, P.; Dennison, C.; Farhi, D.; Fischer, Q.; Hashme, S.; Hesse, C.; Józefowicz, R.; Gray, S.; Olsson, C.; Pachocki, J. W.; Petrov, M.; de Oliveira Pinto, H. P.; Raiman, J.; Salimans, T.; Schlatter, J.; Schneider, J.; Sidor, S.; Sutskever, I.; Tang, J.; Wolski, F.; and Zhang, S. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *ArXiv*, abs/1912.06680.

Bhandari, J.; Russo, D.; and Singal, R. 2018. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, 1691–1692. PMLR.

Bowling, M.; and Veloso, M. 2001. Rational and convergent learning in stochastic games. In *International joint conference on artificial intelligence*, volume 17, 1021–1026. Citeseer.

Cai, Y.; and Daskalakis, C. 2011. On minmax theorems for multiplayer games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete algorithms*, 217–234. SIAM.

Cai, Y.; Luo, H.; Wei, C.-Y.; and Zheng, W. 2024. Uncoupled and Convergent Learning in Two-Player Zero-Sum Markov Games with Bandit Feedback. *Advances in Neural Information Processing Systems*, 36.

Cen, S.; Wei, Y.; and Chi, Y. 2021. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34: 27952–27964.

Chen, X.; Deng, X.; and Teng, S.-H. 2009. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM (JACM)*, 56(3): 1–57.

Chen, X.; and Zhao, L. 2024. Finite-time analysis of single-timescale actor-critic. *Advances in Neural Information Processing Systems*, 36.

Chen, Z.; Ma, S.; and Zhou, Y. 2021. Sample efficient stochastic policy extragradient algorithm for zero-sum Markov game. In *International Conference on Learning Representations*.

Chen, Z.; Zhang, K.; Mazumdar, E.; Ozdaglar, A.; and Wierman, A. 2023. A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games. *arXiv preprint arXiv:2303.03100*.

Daskalakis, C.; Foster, D. J.; and Golowich, N. 2020. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33: 5527–5540.

Daskalakis, C.; Goldberg, P. W.; and Papadimitriou, C. H. 2009. The complexity of computing a Nash equilibrium. *Communications of the ACM*, 52(2): 89–97.

Durrett, R. 2019. *Probability: theory and examples*, volume 49. Cambridge university press.

Fiechter, C.-N. 1994. Efficient reinforcement learning. In *Proceedings of the seventh annual conference on Computational learning theory*, 88–97.

Filar, J. A.; and Tolwinski, B. 1991. On the algorithm of Pollatschek and Avi-ltzhak. In *Stochastic Games And Related Topics: In Honor of Professor LS Shapley*, 59–70. Springer.

Freund, Y.; and Schapire, R. E. 1999. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2): 79–103.

Gao, B.; and Pavel, L. 2017. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*.

Hofbauer, J.; and Hopkins, E. 2005. Learning in perturbed asymmetric games. *Games and Economic Behavior*, 52(1): 133–152.

Hofbauer, J.; and Sorin, S. 2006. Best response dynamics for continuous zero-sum games. *Discrete and Continuous Dynamical Systems Series B*, 6(1): 215.

Hoffman, A. J.; and Karp, R. M. 1966. On nonterminating stochastic games. *Management Science*, 12(5): 359–370.

Jerrum, M. 2003. *Counting, sampling and integrating: algorithms and complexity*. Springer Science & Business Media.

Khodadadian, S.; Doan, T. T.; Romberg, J.; and Maguluri, S. T. 2022. Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*.

Konda, V. R.; and Borkar, V. S. 1999. Actor-critic–type learning algorithms for Markov decision processes. *SIAM Journal on control and Optimization*, 38(1): 94–123.

Kumar, H.; Koppel, A.; and Ribeiro, A. 2023. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, 1–35.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, 157–163. Elsevier.

Liu, Q.; Wang, Y.; and Jin, C. 2022. Learning markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *International Conference on Machine Learning*, 14036–14053. PMLR.

Mei, J.; Xiao, C.; Szepesvari, C.; and Schuurmans, D. 2020. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, 6820–6829. PMLR.

Montenegro, R.; Tetali, P.; et al. 2006. Mathematical aspects of mixing times in Markov chains. *Foundations and Trends® in Theoretical Computer Science*, 1(3): 237–354.

Muthukumar, V.; Phade, S.; and Sahai, A. 2020. On the Impossibility of Convergence of Mixed Strategies with No Regret Learning. *arXiv preprint arXiv:2012.02125*.

Olshevsky, A.; and Gharesifard, B. 2023. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2): 980–1007.

Ouhamma, R.; and Kamgarpour, M. 2023. Learning nash equilibria in zero-sum markov games: A single timescale algorithm under weak reachability. *arXiv preprint arXiv:2312.08008*.

Pollatschek, M.; and Avi-Itzhak, B. 1969. Algorithms for stochastic games with geometrical interpretation. *Management Science*, 15(7): 399–415.

Sayin, M.; Zhang, K.; Leslie, D.; Basar, T.; and Ozdaglar, A. 2021. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems*, 34: 18320–18334.

Shapley, L. S. 1953. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100.

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

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52: 479–487.

Van Der Wal, J. 1978. Discounted Markov games: Generalized policy iteration method. *Journal of Optimization Theory and Applications*, 25: 125–138.

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

Wei, C.-Y.; Hong, Y.-T.; and Lu, C.-J. 2017. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30.

Wei, C.-Y.; Lee, C.-W.; Zhang, M.; and Luo, H. 2021. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on learning theory*, 4259–4299. PMLR.

Wu, Y. F.; Zhang, W.; Xu, P.; and Gu, Q. 2020. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33: 17617–17628.

Xu, P.; Gao, F.; and Gu, Q. 2020. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, 541–551. PMLR.

Xu, T.; Wang, Z.; and Liang, Y. 2020. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33: 4358–4369.

Zhang, R. C.; Ren, Z.; and Li, N. 2022. Gradient play in stochastic games: Stationary points and local geometry. *IFAC-PapersOnLine*, 55(30): 73–78.

Zhang, Y.; Qu, G.; Xu, P.; Lin, Y.; Chen, Z.; and Wierman, A. 2023. Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1): 1–51.

Zhao, Y.; Tian, Y.; Lee, J.; and Du, S. 2022. Provably efficient policy optimization for two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, 2736–2761. PMLR.

Zimmert, J.; and Seldin, Y. 2021. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1): 1310–1358.