FASTER LAST-ITERATE CONVERGENCE OF POLICY OPTIMIZATION IN ZERO-SUM MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-Agent Reinforcement Learning (MARL)—where multiple agents learn to interact in a shared dynamic environment—permeates across a wide range of critical applications. While there has been substantial progress on understanding the global convergence of policy optimization methods in single-agent RL, designing and analysis of efficient policy optimization algorithms in the MARL setting present significant challenges, which unfortunately, remain highly inadequately addressed by existing theory. In this paper, we focus on the most basic setting of competitive multi-agent RL, namely two-player zero-sum Markov games, and study equilibrium finding algorithms in both the infinite-horizon discounted setting and the finite-horizon episodic setting. We propose a single-loop policy optimization method with symmetric updates from both agents, where the policy is updated via the entropy-regularized optimistic multiplicative weights update (OMWU) method and the value is updated on a slower timescale. We show that, in the full-information tabular setting, the proposed method achieves a finite-time last-iterate linear convergence to the quantal response equilibrium of the regularized problem, which translates to a sublinear last-iterate convergence to the Nash equilibrium by controlling the amount of regularization. Our convergence results improve upon the best known iteration complexities, and lead to a better understanding of policy optimization in competitive Markov games.

Keywords: zero-sum Markov game, entropy regularization, policy optimization, global convergence, multiplicative updates

1 INTRODUCTION

Policy optimization methods (Williams, 1992; Sutton et al., 2000; Kakade, 2002; Peters and Schaal, 2008; Konda and Tsitsiklis, 2000), which cast sequential decision making as value maximization problems with regards to (parameterized) policies, have been instrumental in enabling recent successes of reinforcement learning (RL). See e.g., Schulman et al. (2015; 2017); Silver et al. (2016). Despite its empirical popularity, the theoretical underpinnings of policy optimization methods remain elusive until very recently. For single-agent RL problems, a flurry of recent works has made substantial progress on understanding the global convergence of policy optimization methods under the framework of Markov Decision Processes (MDP) (Agarwal et al., 2020; Bhandari and Russo, 2019; Mei et al., 2020; Cen et al., 2021a; Lan, 2022; Bhandari and Russo, 2020; Zhan et al., 2021; Khodadadian et al., 2021; Xiao, 2022). Despite the nonconcave nature of value maximization, (natural) policy gradient methods are shown to achieve global convergence at a sublinear rate (Agarwal et al., 2020; Mei et al., 2020) or even a linear rate in the presence of regularization (Mei et al., 2020; Cen et al., 2021a; Lan, 2022; Zhan et al., 2021) when the learning rate is constant.

Moving beyond single-agent RL, Multi-Agent Reinforcement Learning (MARL) is the next frontier—where multiple agents learn to interact in a shared dynamic environment—permeating across critical applications such as multi-agent networked systems, autonomous vehicles, robotics, and so on. Designing and analysis of efficient policy optimization algorithms in the MARL setting present significant challenges and new desiderata, which unfortunately, remain highly inadequately addressed by existing theory.

1.1 POLICY OPTIMIZATION FOR COMPETITIVE RL

In this work, we focus on one of the most basic settings of competitive multi-agent RL, namely two-player zero-sum Markov games (Shapley, 1953), and study equilibrium finding algorithms in both the infinite-horizon discounted setting and the finite-horizon episodic setting. In particular, our designs gravitate around algorithms that are *single-loop*, *symmetric*, with *finite-time last-iterate* convergence to the Nash Equilibrium (NE) or Quantal Response Equilibrium (QRE) under bounded rationality, two prevalent solution concepts in game theory. These design principles naturally come up as a result of pursuing simple yet efficient algorithms: *single-loop* updates preclude sophisticated interleaving of rounds between agents; *symmetric* updates ensure no agent will compromise its rewards in the learning process, which can be otherwise exploited by a faster-updating opponent; in addition, asymmetric updates typically lead to one-sided convergence, i.e., only one of the agents is guaranteed to converge to the minimax equilibrium in a non-asymptotic manner, which is less desirable; moreover, *last-iterate convergence* guarantee absolves the need for agents to switch between learning and deployment; last but not least, it is desirable to converge as fast as possible, where the iteration complexities are *non-asymptotic* with clear dependence on salient problem parameters.

Substantial algorithmic developments have been made for finding equilibria in two-player zero-sum Markov games, where Dynamical Programming (DP) techniques have long been used as a fundamental building block, leading to prototypical iterative schemes such as Value Iteration (VI) (Shapley, 1953) and Policy Iteration (PI) (Van Der Wal, 1978; Patek and Bertsekas, 1999). Different from their single-agent counterparts, these methods require solving a two-player zero-sum matrix game for every state per iteration. A considerable number of recent works (Zhao et al., 2022; Alacaoglu et al., 2022; Cen et al., 2021b; Chen et al., 2021a) are based on these DP iterations, by plugging in various (gradient-based) solvers of two-player zero-sum matrix games. However, these methods are inherently nested-loop, which are less convenient to implement. In addition, PI-based methods are asymmetric and come with only one-sided convergence guarantees (Patek and Bertsekas, 1999; Zhao et al., 2022; Alacaoglu et al., 2022).

Going beyond nested-loop algorithms, single-loop policy gradient methods have been proposed recently for solving two-player zero-sum Markov games. Here, we are interested in finding an ϵ-optimal NE or QRE in terms of the duality gap, i.e. the difference in the value functions when either of the agents deviates from the solution policy.

- For the infinite-horizon discounted setting, Daskalakis et al. (2020) demonstrated that the independent policy gradient method, with direct parameterization and asymmetric learning rates, finds an ϵ-optimal NE within a polynomial number of iterations. Zeng et al. (2022) improved over this rate using an entropy-regularized policy gradient method with softmax parameterization and asymmetric learning rates. On the other end, Wei et al. (2021) proposed an optimistic gradient descent ascent (OGDA) method (Rakhlin and Sridharan, 2013) with direct parameterization and symmetric learning rates,[1] which achieves a last-iterate convergence at a rather pessimistic iteration complexity.
- For the finite-horizon episodic setting, Zhang et al. (2022); Yang and Ma (2022) showed that the weighted average-iterate of the optimistic Follow-The-Regularized-Leader (FTRL) method, when combined with slow critic updates, finds an ϵ-optimal NE in a polynomial number of iterations.

A more complete summary of prior results can be found in Table 1 and Table 2. In brief, while there have been encouraging progresses in developing computationally efficient policy gradient methods for solving zero-sum Markov games, achieving fast finite-time last-iterate convergence with single-loop and symmetric update rules remains a challenging goal.

1.2 OUR CONTRIBUTIONS

Motivated by the positive role of entropy regularization in enabling faster convergence of policy optimization in single-agent RL (Cen et al., 2021a; Lan, 2022) and two-player zero-sum games (Cen

[1] To be precise, Wei et al. (2021) proved the average-iterate convergence of the duality gap, as well as the last-iterate convergence of the policy in terms of the Euclidean distance to the set of NEs, where it is possible to translate the latter last-iterate convergence to the duality gap (see Appendix. F). The resulting iteration complexity, however, is much worse than that of the average-iterate convergence in terms of the duality gap, with a problem-dependent constant that can scale pessimistically with salient problem parameters.

et al., 2021b), we propose a single-loop policy optimization algorithm for two-player zero-sum Markov games in both the infinite-horizon and finite-horizon settings. The proposed algorithm follows the style of actor-critic (Konda and Tsitsiklis, 2000), with the actor updating the policy via the entropy-regularized optimistic multiplicative weights update (OMWU) method (Cen et al., 2021b) and the critic updating the value function on a slower timescale. Both agents execute multiplicative and symmetric policy updates, where the learning rates are carefully selected to ensure a fast last-iterate convergence. In both the infinite-horizon and finite-horizon settings, we prove that the last iterate of the proposed method learns the optimal value function and converges at a linear rate to the unique QRE of the entropy-regularized Markov game, which can be further translated into finding the NE by setting the regularization sufficiently small.

- For the infinite-horizon discounted setting, the last iterate of our method takes at most

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^4\tau}\log\frac{1}{\epsilon}\right)$$

  iterations for finding an ϵ-optimal QRE under entropy regularization, where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies. Here, $|\mathcal{S}|$ is the size of the state space, γ is the discount factor, and τ is the regularization parameter. Moreover, this implies the last-iterate convergence with an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^5\epsilon}\right)$$

  for finding an ϵ-optimal NE.

- For the finite-horizon episodic setting, the last iterate of our method takes at most

$$\widetilde{\mathcal{O}}\left(\frac{H^2}{\tau}\log\frac{1}{\epsilon}\right)$$

  iterations for finding an ϵ-optimal QRE under entropy regularization, where H is the horizon length. Similarly, this implies the last-iterate convergence with an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\frac{H^3}{\epsilon}\right)$$

  for finding an ϵ-optimal NE.

Detailed comparisons between the proposed method and prior arts are provided in Table 1 and Table 2. To the best of our knowledge, this work presents the first method that is simultaneously single-loop, symmetric, and achieves fast finite-time last-iterate convergence in terms of the duality gap in both infinite-horizon and finite-horizon settings. From a technical perspective, the infinite-horizon discounted setting is in particular challenging, where ours is the first single-loop algorithm that guarantees an iteration complexity of $\widetilde{\mathcal{O}}(1/\epsilon)$ for last-iterate convergence in terms of the duality gap, with clear and improved dependencies on other problem parameters in the meantime. In contrast, several existing works introduce additional problem-dependent constants (Daskalakis et al., 2020; Wei et al., 2021; Zeng et al., 2022) in the iteration complexity, which can scale rather pessimistically—sometimes even exponentially—with problem dimensions (Li et al., 2021).

Our technical developments require novel ingredients that deviate from prior tools such as error propagation analysis for Bellman operators (Perolat et al., 2015; Patek and Bertsekas, 1999) from a dynamic programming perspective, as well as the gradient dominance condition (Daskalakis et al., 2020; Zeng et al., 2022) from a policy optimization perspective. Importantly, at the core of our analysis lies a carefully-designed one-step error contraction bound for policy learning, together with a set of recursive error bounds for value learning, all of which tailored to the non-Euclidean OMWU update rules that have not been well studied in the setting of Markov games.

1.3 Related works

Learning in two-player zero-sum matrix games. Freund and Schapire (1999) showed that the average iterate of Multiplicative Weight Update (MWU) method converges to NE at a rate of $\mathcal{O}(1/\sqrt{T})$, which in principle holds for many other no-regret algorithms as well. Daskalakis

Solution type	Reference	Iteration complexity	Single loop	Symmetric	Last-iterate convergence
ϵ-NE	**PI-based Methods** **Zhao et al. (2022)** **Alacaoglu et al. (2022)**	$\widetilde{\mathcal{O}}\left(\frac{\lVert 1/\rho\rVert_\infty}{(1-\gamma)^3\epsilon}\right)^*$	✗	✗	✓
	VI-based Methods **Cen et al. (2021b)** **Chen et al. (2021a)**	$\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3\epsilon}\right)$	✗	✓	✓
	Daskalakis et al. (2020)	Polynomial*	✓	✗	✗
	Zeng et al. (2022)	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^2\lVert 1/\rho\rVert_\infty^5}{(1-\gamma)^{14}c^4\epsilon^3}\right)^*$	✓	✗	✓
	Wei et al. (2021)	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^3}{(1-\gamma)^9\epsilon^2}\right)$	✓	✓	✗
		$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^5(\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert)^{1/2}}{(1-\gamma)^{16}c^4\epsilon^2}\right)$	✓	✓	✓
	This Work	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert}{(1-\gamma)^5\epsilon}\right)$	✓	✓	✓
ϵ-QRE	**VI-based Methods** **Cen et al. (2021b)**	$\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3}\log^2\frac{1}{\epsilon}\right)$	✗	✓	✓
	Zeng et al. (2022)	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^2\lVert 1/\rho\rVert_\infty^5}{(1-\gamma)^{11}c^4\tau^3}\log\frac{1}{\epsilon}\right)^*$	✓	✗	✓
	This Work	$\widetilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert}{(1-\gamma)^4\tau}\log\frac{1}{\epsilon}\right)$	✓	✓	✓

Table 1: Comparison of policy optimization methods for finding an ϵ-optimal NE or QRE of two-player zero-sum discounted Markov games in terms of the duality gap. Note that $*$ implies one-sided convergence, i.e., only one of the agents is guaranteed to achieve finite-time convergence to the equilibrium. Here, $c > 0$ refers to some problem-dependent constant. For simplicity and a fair comparison, we replace various notions of concentrability coefficient and distribution mismatch coefficient with a crude upper bound $\lVert 1/\rho\rVert_\infty$, where ρ is the initial state distribution.

Solution type	Reference	Iteration complexity	Single loop	Symmetric	Last-iterate convergence
ϵ-NE	**Zhang et al. (2022)** **OFTRL**	$\widetilde{\mathcal{O}}\left(\frac{H^{28/5}}{\epsilon^{6/5}}\right)$	✓	✓	✗
	Zhang et al. (2022) **modified OFTRL**	$\widetilde{\mathcal{O}}\left(\frac{H^4}{\epsilon}\right)$	✓	✓	✗
	Yang and Ma (2022) **OFTRL**	$\widetilde{\mathcal{O}}\left(\frac{H^5}{\epsilon}\right)$	✓	✓	✗
	This Work	$\widetilde{\mathcal{O}}\left(\frac{H^3}{\epsilon}\right)$	✓	✓	✓
ϵ-QRE	**This Work**	$\widetilde{\mathcal{O}}\left(\frac{H^2}{\tau}\log\frac{1}{\epsilon}\right)$	✓	✓	✓

Table 2: Comparison of policy optimization methods for finding an ϵ-optimal NE or QRE of two-player zero-sum episodic Markov games in terms of the duality gap.

et al. (2011) deployed the excessive gap technique of Nesterov and improved the convergence rate to $\mathcal{O}(1/T)$, which is achieved later by (Rakhlin and Sridharan, 2013) with a simple modification of MWU method, named Optimistic Mirror Descent (OMD) or more commonly, OMWU. Moving beyond average-iterate convergence, Bailey and Piliouras (2018) demonstrated that MWU updates, despite converging in an ergodic manner, diverge from the equilibrium. Daskalakis and Panageas (2018); Wei et al. (2021) explored the last-iterate convergence guarantee of OMWU, assuming uniqueness of NE. Cen et al. (2021b) established linear last-iterate convergence of entropy-regularized OMWU without uniqueness assumption. Sokota et al. (2022) showed that optimistic update is not necessary for achieving linear last-iterate convergence in the presence of regularization, albeit with a more strict restriction on the step size.

Learning in two-player zero-sum Markov games. In addition to the aforementioned works on policy optimization methods (policy-based methods) for two-player zero-sum Markov games (cf. Table 1 and Table 2), a growing body of works have developed model-based methods (Liu et al., 2021; Zhang et al., 2020; Li et al., 2022) and value-based methods (Bai and Jin, 2020; Bai et al., 2020; Chen et al., 2021b; Jin et al., 2021; Sayin et al., 2021; Xie et al., 2020), with a primary focus on learning NE in a sample-efficient manner. Our work, together with prior literatures on policy optimization, focuses instead on learning NE in a computation-efficient manner assuming full-information.

Entropy regularization in RL and games. Entropy regularization is a popular algorithmic idea in RL (Williams and Peng, 1991) that promotes exploration of the policy. A recent line of works (Mei et al., 2020; Cen et al., 2021a; Lan, 2022; Zhan et al., 2021) demonstrated that incorporating entropy regularization provably accelerates policy optimization in single-agent MDPs by enabling fast linear convergence. While the positive role of entropy regularization is also verified in various game-theoretic settings, e.g., two-player zero-sum matrix games (Cen et al., 2021b), zero-sum polymatrix games (Leonardos et al., 2021), and potential games (Cen et al., 2022), it remains highly unexplored the interplay between entropy regularization and policy optimization in Markov games with only a few exceptions (Zeng et al., 2022).

1.4 Notations

We denote the probability simplex over a set $\mathcal{A}$ by $\Delta(\mathcal{A})$. We use bracket with subscript to index the entries of a vector or matrix, e.g., $[x]_a$ for a-th element of a vector x, or simply $x(a)$ when it is clear from the context. Given two distributions $x, y \in \Delta(\mathcal{A})$, the Kullback-Leibler (KL) divergence from y to x is denoted by $\mathsf{KL}\big(x \,\|\, y\big) = \sum_{a\in\mathcal{A}} x(a)(\log x(a) - \log y(a))$. Finally, we denote by $\|A\|_\infty$ the maximum entrywise absolute value of a matrix A, i.e., $\|A\|_\infty = \max_{i,j} |A_{i,j}|$.

2 Algorithm and Theory: The Infinite-Horizon Setting

2.1 Problem Formulation

Two-player zero-sum discounted Markov game. A two-player zero-sum discounted Markov game is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, \gamma)$, with finite state space $\mathcal{S}$, finite action spaces of the two players $\mathcal{A}$ and $\mathcal{B}$, reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$, transition probability kernel $P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ and discount factor $0 \le \gamma < 1$. The action selection rule of the max player (resp. the min player) is represented by $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ (resp. $\nu : \mathcal{S} \to \Delta(\mathcal{B})$), where the probability of selecting action $a \in \mathcal{A}$ (resp. $b \in \mathcal{B}$) in state $s \in \mathcal{S}$ is specified by $\mu(a|s)$ (resp. $\nu(b|s)$). The probability of transitioning from state s to a new state s' upon selecting the action pair $(a, b) \in \mathcal{A}, \mathcal{B}$ is given by $P(s'|s, a, b)$.

Value function and Q-function. For a given policy pair μ, ν, the state value of $s \in \mathcal{S}$ is evaluated by the expected discounted sum of rewards with initial state $s_0 = s$:

$$\forall s \in \mathcal{S}: \qquad V^{\mu,\nu}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \big| s_0 = s\right], \tag{1}$$

the quantity the max player seeks to maximize while the min player seeks to minimize. Here, the trajectory $(s_0, a_0, b_0, s_1, \cdots)$ is generated according to $a_t \sim \mu(\cdot|s_t)$, $b_t \sim \nu(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t, b_t)$. Similarly, the Q-function $Q^{\mu,\nu}(s, a, b)$ evaluates the expected discounted cumulative reward with initial state s and initial action pair (a, b):

$$\forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}: \qquad Q^{\mu,\nu}(s, a, b) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \big| s_0 = s, a_0 = a, b_0 = b\right]. \tag{2}$$

For notation simplicity, we denote by $Q^{\mu,\nu}(s) \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ the matrix $[Q^{\mu,\nu}(s, a, b)]_{(a,b)\in\mathcal{A}\times\mathcal{B}}$, so that

$$\forall s \in \mathcal{S}: \qquad V^{\mu,\nu}(s) = \mu(s)^\top Q^{\mu,\nu}(s)\nu(s).$$

Shapley (1953) proved the existence of a policy pair $(\mu^\star, \nu^\star)$ that solves the min-max problem

$$\max_{\mu} \min_{\nu} V^{\mu,\nu}(s)$$

for all $s \in \mathcal{S}$ simultaneously, and that the mini-max value is unique. A set of such optimal policy pair $(\mu^\star, \nu^\star)$ is called the Nash equilibrium (NE) to the Markov game.

Entropy regularized two-player zero-sum Markov game. Entropy regularization is shown to provably accelerate convergence in single-agent RL (Geist et al., 2019; Mei et al., 2020; Cen et al., 2021a) and facilitate the analysis in two-player zero-sum matrix games (Cen et al., 2021b) as well as Markov games (Cen et al., 2021b; Zeng et al., 2022). The entropy-regularized value function $V_\tau^{\mu,\nu}(s)$ is defined as

$$\forall s \in \mathcal{S}: \qquad V_\tau^{\mu,\nu}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r(s_t, a_t, b_t) - \tau \log \mu(a_t|s_t) + \tau \log \nu(b_t|s_t) \Big) \Big| s_0 = s \right], \tag{3}$$

where $\tau \geq 0$ is the regularization parameter. Similarly, the regularized Q-function $Q_\tau^{\mu,\nu}$ is given by

$$\forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}: \qquad Q_\tau^{\mu,\nu}(s) = r(s,a,b) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a,b)} \left[V_\tau^{\mu,\nu}(s') \right]. \tag{4}$$

It is known that (Cen et al., 2021b) there exists a unique pair of policy $(\mu_\tau^\star, \nu_\tau^\star)$ that solves the min-max entropy-regularized problem

$$\max_\mu \min_\nu \; V_\tau^{\mu,\nu}(s), \tag{5a}$$

or equivalently

$$\max_\mu \min_\nu \; \mu(s)^\top Q_\tau^{\mu,\nu}(s) \nu(s) + \tau \mathcal{H}\big(\mu(s)\big) - \tau \mathcal{H}\big(\nu(s)\big) \tag{5b}$$

for all $s \in \mathcal{S}$, and we call $(\mu_\tau^\star, \nu_\tau^\star)$ the quantal response equilibrium (QRE) (McKelvey and Palfrey, 1995) to the entropy-regularized Markov game. We denote the associated regularized value function and Q-function by

$$V_\tau^\star(s) = V_\tau^{\mu_\tau^\star, \nu_\tau^\star}(s) \quad \text{and} \quad Q_\tau^\star(s,a,b) = Q_\tau^{\mu_\tau^\star, \nu_\tau^\star}(s,a,b).$$

Goal. We seek to find an ϵ-optimal QRE or ϵ-QRE (resp. ϵ-optimal NE or ϵ-NE) $\zeta = (\mu, \nu)$ which satisfies

$$\max_{s \in \mathcal{S}, \mu', \nu'} \Big(V_\tau^{\mu',\nu}(s) - V_\tau^{\mu,\nu'}(s) \Big) \leq \epsilon \tag{6}$$

(resp. $\max_{s \in \mathcal{S}, \mu', \nu'} \Big(V^{\mu',\nu}(s) - V^{\mu,\nu'}(s) \Big) \leq \epsilon$) in a computationally efficient manner. In truth, the solution concept of ϵ-QRE provides an approximation of ϵ-NE with appropriate choice of the regularization parameter τ. Basic calculations tell us that

$$\begin{aligned} V^{\mu',\nu}(s) - V^{\mu,\nu'}(s) &= \big(V_\tau^{\mu',\nu}(s) - V_\tau^{\mu,\nu'}(s)\big) + \big(V^{\mu',\nu}(s) - V_\tau^{\mu',\nu}(s)\big) - \big(V^{\mu,\nu'}(s) - V_\tau^{\mu,\nu'}(s)\big) \\ &\leq V_\tau^{\mu',\nu}(s) - V_\tau^{\mu,\nu'}(s) + \frac{\tau(\log|\mathcal{A}| + \log|\mathcal{B}|)}{1-\gamma}, \end{aligned}$$

which guarantees that an $\epsilon/2$-QRE is an ϵ-NE as long as $\tau \leq \frac{(1-\gamma)\epsilon}{2(\log|\mathcal{A}| + \log|\mathcal{B}|)}$. For technical convenience, we assume $\tau \leq \frac{1}{\max\{1, \log|\mathcal{A}| + \log|\mathcal{B}|\}}$ throughout the paper.

Additional notation. For notation convenience, we denote by ζ the concatenation of a policy pair μ and ν, i.e., $\zeta = (\mu, \nu)$. The QRE to the regularized problem is denoted by $\zeta_\tau^\star = (\mu_\tau^\star, \nu_\tau^\star)$. We use shorthand notation $\mu(s)$ and $\nu(s)$ to denote $\mu(\cdot|s)$ and $\nu(\cdot|s)$. In addition, we write $\mathsf{KL}\big(\mu(s) \,\|\, \mu'(s)\big)$ and $\mathsf{KL}\big(\nu(s) \,\|\, \nu'(s)\big)$ as $\mathsf{KL}_s\big(\mu \,\|\, \mu'\big)$ and $\mathsf{KL}_s\big(\nu \,\|\, \nu'\big)$, and let

$$\mathsf{KL}_s\big(\zeta \,\|\, \zeta'\big) = \mathsf{KL}_s\big(\mu \,\|\, \mu'\big) + \mathsf{KL}_s\big(\nu \,\|\, \nu'\big).$$

By a slight abuse of notation, $\mathsf{KL}_\rho\big(\zeta \,\|\, \zeta'\big)$ denotes $\mathbb{E}_{s \sim \rho}\left[\mathsf{KL}_s\big(\zeta \,\|\, \zeta'\big)\right]$ for $\rho \in \Delta(\mathcal{S})$.

2.2 Single-loop algorithm design

In this section, we propose a single-loop policy optimization algorithm for finding the QRE of the entropy-regularized Markov game, which is generalized from the entropy-regularized OMWU method (Cen et al., 2021b) for solving entropy-regularized matrix games, with a careful orchestrating of the policy update and the value update.

Algorithm 1: Entropy-regularized OMWU for Discounted Two-player Zero-sum Markov Game

Input: Regularization parameter $\tau > 0$, learning rate for policy update $\eta > 0$, learning rate for value update $\{\alpha_t\}_{t=1}^{\infty}$.

Initialization: Set $\mu^{(0)}, \bar{\mu}^{(0)}, \nu^{(0)}$ and $\bar{\nu}^{(0)}$ as uniform policies; and set

$$Q^{(0)} = 0, \quad V^{(0)} = \tau(\log|\mathcal{A}| - \log|\mathcal{B}|).$$

for $t = 0, 1, \cdots$ **do**

  for *all* $s \in \mathcal{S}$ **do in parallel**

    When $t \geq 1$, update policy pair $\zeta^{(t)}(s)$ as:

$$\begin{cases} \mu^{(t)}(a|s) \propto \mu^{(t-1)}(a|s)^{1-\eta\tau} \exp(\eta[Q^{(t)}(s)\bar{\nu}^{(t)}(s)]_a) \\ \nu^{(t)}(b|s) \propto \nu^{(t-1)}(b|s)^{1-\eta\tau} \exp(-\eta[Q^{(t)}(s)^\top \bar{\mu}^{(t)}(s)]_b) \end{cases}. \tag{9a}$$

    Update policy pair $\bar{\zeta}^{(t+1)}(s)$ as:

$$\begin{cases} \bar{\mu}^{(t+1)}(a|s) \propto \mu^{(t)}(a|s)^{1-\eta\tau} \exp(\eta[Q^{(t)}(s)\bar{\nu}^{(t)}(s)]_a) \\ \bar{\nu}^{(t+1)}(b|s) \propto \nu^{(t)}(b|s)^{1-\eta\tau} \exp(-\eta[Q^{(t)}(s)^\top \bar{\mu}^{(t)}(s)]_b) \end{cases}. \tag{9b}$$

    Update $Q^{(t+1)}(s)$ and $V^{(t+1)}(s)$ as

$$\begin{cases} Q^{(t+1)}(s,a,b) &= r(s,a,b) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[V^{(t)}(s')\right] \\ V^{(t+1)}(s) &= (1-\alpha_{t+1})V^{(t)}(s) \\ & \quad +\alpha_{t+1}\left[\bar{\mu}^{(t+1)}(s)^\top Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s) + \tau\mathcal{H}\big(\bar{\mu}^{(t+1)}(s)\big) - \tau\mathcal{H}\big(\bar{\nu}^{(t+1)}(s)\big)\right] \end{cases}. \tag{10}$$

Review: entropy-regularized OMWU for two-player zero-sum matrix games. We briefly review the algorithm design of entropy-regularized OMWU method for two-player zero-sum matrix game (Cen et al., 2021b). The problem of interest can be described as

$$\max_{\mu\in\Delta(\mathcal{A})} \min_{\nu\in\Delta(\mathcal{B})} \mu^\top A\nu + \tau\mathcal{H}(\mu) - \tau\mathcal{H}(\nu), \tag{7}$$

where $A \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ is the payoff matrix of the game. The update rule of entropy-regularized OMWU with learning rate $\eta > 0$ is defined as follows: $\forall a \in \mathcal{A}, b \in \mathcal{B}$,

$$\begin{cases} \mu^{(t)}(a) \propto \mu^{(t-1)}(a)^{1-\eta\tau} \exp(\eta[A\bar{\nu}^{(t)}]_a) \\ \nu^{(t)}(b) \propto \nu^{(t-1)}(b)^{1-\eta\tau} \exp(-\eta[A^\top\bar{\mu}^{(t)}]_b) \end{cases}, \tag{8a}$$

$$\begin{cases} \bar{\mu}^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[A\bar{\nu}^{(t)}]_a) \\ \bar{\nu}^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top\bar{\mu}^{(t)}]_b) \end{cases}. \tag{8b}$$

We remark that the update rule can be alternatively motivated from the perspective of natural policy gradient (Kakade, 2002; Cen et al., 2021a) or mirror descent (Lan, 2022; Zhan et al., 2021) with optimistic updates. In particular, the midpoint $(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})$ serves as a prediction of $(\mu^{(t+1)}, \nu^{(t+1)})$ by running one step of mirror descent. Cen et al. (2021b) established that the last iterate of entropy-regularized OMWU converges to the QRE of the matrix game (7) at a linear rate $(1-\eta\tau)^t$, as long as the step size η is no larger than $\min\left\{\frac{1}{2\|A\|_\infty+2\tau}, \frac{1}{4\|A\|_\infty}\right\}$.

Single-loop algorithm for two-player zero-sum Markov games. In view of the similarity in the problem formulations of (5b) and (7), it is tempting to apply the aforementioned method to the Markov game in a state-wise manner, where the Q-function assumes the role of the payoff matrix. It is worth noting, however, that Q-function depends on the policy pair $\zeta = (\mu, \nu)$ and is hence changing concurrently with the update of the policy pair. We take inspiration from Wei et al. (2021) and equip the entropy-regularized OMWU method with the following update rule that iteratively approximates the value function in an actor-critic fashion:

$$Q^{(t+1)}(s,a,b) = r(s,a,b) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[V^{(t)}(s')\right],$$

where $V^{(t+1)}$ is updated as a convex combination of the previous $V^{(t)}$ and the regularized game value induced by $Q^{(t+1)}$ as well as the policy pair $\bar{\zeta}^{(t+1)} = (\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})$:

$$\begin{aligned} V^{(t+1)}(s) &= (1-\alpha_{t+1})V^{(t)}(s) \\ &\quad + \alpha_{t+1}\big[\bar{\mu}^{(t+1)}(s)^\top Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s) + \tau\mathcal{H}\big(\bar{\mu}^{(t+1)}(s)\big) - \tau\mathcal{H}\big(\bar{\nu}^{(t+1)}(s)\big)\big]. \end{aligned} \tag{11}$$

The update of V becomes more conservative with a smaller learning rate α_t, hence stabilizing the update of policies. However, setting α_t too small slows down the convergence of V to $V_\tau^\star$. A key novelty—suggested by our analysis—is the choice of the constant learning rates $\alpha := \alpha_t = \eta\tau$ which updates at a slower timescale than the policy due to $\tau < 1$. This is in sharp contrast to the vanishing sequence $\alpha_t = \frac{2/(1-\gamma)+1}{2/(1-\gamma)+t}$ adopted in Wei et al. (2021), which is essential in their analysis but inevitably leads to a much slower convergence. We summarize the detailed procedure in Algorithm 1. Last but not least, it is worth noting that the proposed method access the reward via "first-order information", i.e., either agent can only update its policy with the marginalized value function $Q(s)\nu(s)$ or $Q(s)^\top\mu(s)$. Update rules of this kind are instrumental in breaking the curse of multi-agents in the sample complexity when adopting sample-based estimates in (10), as we only need to estimate the marginalized Q-function rather than its full form (Li et al., 2022; Chen et al., 2021a).

2.3 Theoretical guarantees

Below we present our main results concerning the last-iterate convergence of Algorithm 1 for solving entropy-regularized two-player zero-sum Markov games in the infinite-horizon discounted setting. The proof is postponed to Appendix A.

Theorem 1. *Setting $0 < \eta \le \frac{(1-\gamma)^3}{32000|\mathcal{S}|}$ and $\alpha_t = \eta\tau$, it holds for all $t \ge 0$ that*

$$\max\left\{\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big),\ \frac{1}{2|\mathcal{S}|}\sum_{s\in\mathcal{S}}\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t)}\big),\ \frac{3\eta}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\big\|Q^{(t)}(s) - Q_\tau^\star(s)\big\|_\infty\right\} \le \frac{3000}{(1-\gamma)^2\tau}\Big(1 - \frac{(1-\gamma)\eta\tau}{4}\Big)^t; \tag{12a}$$

and

$$\max_{s\in\mathcal{S},\mu,\nu}\Big(V_\tau^{\mu,\bar{\nu}^{(t)}}(s) - V_\tau^{\bar{\mu}^{(t)},\nu}(s)\Big) \le \frac{6000|\mathcal{S}|}{(1-\gamma)^3\tau}\max\left\{\frac{8}{(1-\gamma)^2\tau}, \frac{1}{\eta}\right\}\Big(1 - \frac{(1-\gamma)\eta\tau}{4}\Big)^t. \tag{12b}$$

Theorem 1 demonstrates that as long as the learning rate η is small enough, the last iterate of Algorithm 1 converges at a linear rate for the entropy-regularized Markov game. Compared with prior literatures investigating on policy optimization, our analysis focuses on the last-iterate convergence of non-Euclidean updates in the presence of entropy regularization, which appears to be the first of its kind. Several remarks are in order, with detailed comparisons in Table 1.

- **Linear convergence to the QRE.** Theorem 1 demonstrates that the last iterate of Algorithm 1 takes at most $\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)\eta\tau}\log\frac{1}{\epsilon}\right)$ iterations to yield an ϵ-optimal policy in terms of the KL divergence to the QRE $\max_{s\in\mathcal{S}}\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t)}\big) \le \epsilon$, the entrywise error of the regularized Q-function $\big\|Q^{(t)} - Q_\tau^\star\big\|_\infty \le \epsilon$, as well as the duality gap $\max_{s\in\mathcal{S},\mu,\nu}\big(V_\tau^{\mu,\bar{\nu}^{(t)}}(s) - V_\tau^{\bar{\mu}^{(t)},\nu}(s)\big) \le \epsilon$ at once. Minimizing the bound over the learning rate η, the proposed method is guaranteed to find an ϵ-QRE within $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^4\tau}\log\frac{1}{\epsilon}\right)$ iterations, which significantly improves upon the one-side convergence rate of Zeng et al. (2022).
- **Last-iterate convergence to ϵ-optimal NE.** By setting $\tau = \frac{(1-\gamma)\epsilon}{2(\log|\mathcal{A}|+\log|\mathcal{B}|)}$, this immediately leads to provable last-iterate convergence to an ϵ-NE, with an iteration complexity of $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^5\epsilon}\right)$, which again outperforms the convergence rate in Wei et al. (2021).

Remark 1. *The learning rate η is constrained to be inverse proportional to $|\mathcal{S}|$, which is for the worst case and can be potentially loosened for problems with a small concentration coefficient. We refer interested readers to Appendix A for details.*

3 Algorithm and Theory: The Episodic Setting

Episodic two-player zero-sum Markov game. An episodic two-player zero-sum Markov game is defined by a tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H\}$, with $\mathcal{S}$ being a finite state space, $\mathcal{A}$ and $\mathcal{B}$ denoting finite action spaces of the two players, and $H > 0$ the horizon length. Every step $h \in [H]$ admits a transition probability kernel $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and reward function $r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$. Furthermore, $\mu = \{\mu_h\}_{h=1}^H$ and $\{\nu_h\}_{h=1}^H$ denote the policies of the two players, where the probability of the max player choosing $a \in \mathcal{A}$ (resp. the min player choosing $b \in \mathcal{B}$) at time h is specified by $\mu_h(a|s)$ (resp. $\nu_h(a|s)$).

Entropy regularized value functions. The value function and Q-function characterize the expected cumulative reward starting from step h by following the policy pair μ, ν. For conciseness, we only present the definition of entropy-regularized value functions below and remark that the their un-regularized counterparts $V_h^{\mu,\nu}$ and $Q_h^{\mu,\nu}$ can be obtained by setting $\tau = 0$. We have

$$V_{h,\tau}^{\mu,\nu}(s) = \mathbb{E}\left[\sum_{h'=h}^{H} \left[r_{h'}(s_{h'}, a_{h'}, b_{h'}) - \tau \log \mu_{h'}(a_{h'}|s_{h'}) + \tau \log \nu_{h'}(b_{h'}|s_{h'})\right] \;\middle|\; s_h = s\right];$$

$$Q_{h,\tau}^{\mu,\nu}(s, a, b) = r_h(s, a, b) + \mathbb{E}_{s' \sim P_h(\cdot|s,a,b)}\left[V_{h+1,\tau}^{\mu,\nu}(s')\right].$$

The solution concept of NE and QRE are defined in a similar manner by focusing on the episodic versions of value functions. We again denote the unique QRE by $\zeta_\tau^\star = (\mu_\tau^\star, \nu_\tau^\star)$.

Proposed method and convergence guarantee It is straightforward to adapt Algorithm 1 to the episodic setting with minimal modifications, with detailed procedure showcased in Algorithm 1 (cf. Appendix B). The analysis, which substantially deviates from the discounted setting, exploits the structure of finite-horizon MDP and time-inhomogeneous policies, enabling a much larger range of learning rates as showed in the following theorem.

Theorem 2. *Setting $0 < \eta \le \frac{1}{8H}$ and $\alpha_t = \eta\tau$, it holds for all $h \in [H]$ and $t \ge T_h := (H-h)T_{\mathsf{start}}$ with $T_{\mathsf{start}} = \lceil \frac{1}{\eta\tau} \log H \rceil$ that*

$$\left\|Q_{h,\tau}^\star - Q_h^{(t)}\right\|_\infty \le (1-\eta\tau)^{t-T_h} t^{H-h}; \tag{13a}$$

$$\max_{s\in\mathcal{S},\mu,\nu} \left(V_{h,\tau}^{\mu,\bar{\nu}^{(t)}}(s) - V_{h,\tau}^{\bar{\mu}^{(t)},\nu}(s)\right) \le 4(1-\eta\tau)^{t-T_h} \max\left\{\frac{8H^2}{\tau}, \frac{1}{\eta}\right\}\left(\frac{8H}{\tau} + 6\eta t^{H-h+1}\right). \tag{13b}$$

Theorem 2 implies that the last iterate of Algorithm 1 takes no more than $\widetilde{\mathcal{O}}\big(HT_{\mathsf{start}} + \frac{H}{\eta\tau}\log\frac{1}{\epsilon}\big) = \widetilde{\mathcal{O}}\big(\frac{H}{\eta\tau}\log\frac{1}{\epsilon}\big)$ iterations for finding an ϵ-QRE. Minimizing the bound over the learning rate η, Algorithm 1 is guaranteed to find an ϵ-QRE in $\widetilde{\mathcal{O}}\left(\frac{H^2}{\tau}\log\frac{1}{\epsilon}\right)$ iterations, which translates into an iteration complexity of $\widetilde{\mathcal{O}}\left(\frac{H^3}{\epsilon}\right)$ for finding an ϵ-NE in terms of the duality gap, i.e., $\max_{s\in\mathcal{S},h\in[H],\mu,\nu}\left(V_h^{\mu,\bar{\nu}^{(t)}}(s) - V_h^{\bar{\mu}^{(t)},\nu}(s)\right) \le \epsilon$, by setting $\tau = \mathcal{O}\Big(\frac{\epsilon}{H(\log|\mathcal{A}|+\log|\mathcal{B}|)}\Big)$.

4 Discussion

This work develops policy optimization methods for zero-sum Markov games that feature single-loop and symmetric updates with provable last-iterate convergence guarantees. Our approach yields better iteration complexities in both infinite-horizon and finite-horizon settings, by adopting entropy regularization and non-Euclidean policy update. Important future directions include investigating whether larger learning rates are possible without knowing problem-dependent information a priori, extending the framework to allow function approximation, and designing sample-efficient implementations of the proposed method.

REFERENCES

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.

A. Alacaoglu, L. Viano, N. He, and V. Cevher. A natural actor-critic framework for zero-sum Markov games. In *International Conference on Machine Learning*, pages 307–366. PMLR, 2022.

Y. Bai and C. Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560. PMLR, 2020.

Y. Bai, C. Jin, and T. Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

J. P. Bailey and G. Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 321–338, 2018.

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

J. Bhandari and D. Russo. A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*, 2020.

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

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

S. Cen, F. Chen, and Y. Chi. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization. In *2022 IEEE 61th Conference on Decision and Control (CDC)*. IEEE, 2022.

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

Z. Chen, D. Zhou, and Q. Gu. Almost optimal algorithms for two-player Markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*, 2021b.

C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv preprint arXiv:1807.04252*, 2018.

C. Daskalakis, A. Deckelbaum, and A. Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

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.

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

M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169, 2019.

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.

S. M. Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

S. Khodadadian, P. R. Jhunjhunwala, S. M. Varma, and S. T. Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer, 2000.

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022.

S. Leonardos, G. Piliouras, and K. Spendlove. Exploration-exploitation in multi-agent competition: convergence with bounded rationality. *Advances in Neural Information Processing Systems*, 34: 26318–26331, 2021.

G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pages 3107–3110. PMLR, 2021.

G. Li, Y. Chi, Y. Wei, and Y. Chen. Minimax-optimal multi-agent RL in zero-sum Markov games with a generative model. *arXiv preprint arXiv:2208.10458*, 2022.

Q. Liu, T. Yu, Y. Bai, and C. Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

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

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.

S. D. Patek and D. P. Bertsekas. Stochastic shortest path games. *SIAM Journal on Control and Optimization*, 37(3):804–824, 1999.

J. Perolat, B. Scherrer, B. Piot, and O. Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pages 1321–1329. PMLR, 2015.

J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

A. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. *arXiv preprint arXiv:1311.1869*, 2013.

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

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

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.

S. Sokota, R. D'Orazio, J. Z. Kolter, N. Loizou, M. Lanctot, I. Mitliagkas, N. Brown, and C. Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. *arXiv preprint arXiv:2206.05825*, 2022.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

J. Van Der Wal. Discounted markov games: Generalized policy iteration method. *Journal of Optimization Theory and Applications*, 25(1):125–138, 1978.

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. In *Conference on learning theory*, pages 4259–4299. PMLR, 2021.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

R. J. Williams and J. Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

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

Q. Xie, Y. Chen, Z. Wang, and Z. Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

Y. Yang and C. Ma. $O(T^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games. *arXiv preprint arXiv:2209.12430*, 2022.

S. Zeng, T. T. Doan, and J. Romberg. Regularized gradient descent ascent for two-player zero-sum Markov games. *Advances in Neural Information Processing Systems*, 35, 2022.

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.

K. Zhang, S. Kakade, T. Basar, and L. Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33, 2020.

R. Zhang, Q. Liu, H. Wang, C. Xiong, N. Li, and Y. Bai. Policy optimization for Markov games: Unified framework and faster convergence. *Advances in Neural Information Processing Systems*, 35, 2022.

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