RegFTRL: Efficient Equilibrium Learning in Two-Player Zero-Sum Games

Zijian Fang
fangzj@mail2.sysu.edu.cn
Sun Yat-sen University
Guangzhou, Guangdong, China

Zongkai Liu
liuzk@mail2.sysu.edu.cn
Sun Yat-sen University
Guangzhou, Guangdong, China

Chao Yu*
yuchao3@mail.sysu.edu.cn
Sun Yat-sen University
Guangzhou, Guangdong, China

Abstract

Recent literature has witnessed a rising interest in learning Nash equilibrium with a guarantee of last-iterate convergence. In this paper, we introduce a novel approach called Regularized Follow-the-Regularized-Leader (RegFTRL), which is an efficient variant of FTRL enriched with an adaptive regularization, for the purpose of learning equilibria in two-player zero-sum games. In the context of normal-form games (NFGs), our proposed RegFTRL algorithm exhibits desirable property of last-iterate linear convergence towards an approximated equilibrium, and converges to an exact Nash equilibrium through adaptive adjustments of the regularization. Moreover, we extend our method to extensive-form games (EFGs) and propose FollowMu, a practical implementation of RegFTRL with a neural network as the function approximator, for model-free learning in sequential non-stationary environments. Finally, empirical results substantiate the theoretical properties of RegFTRL, and demonstrate that FollowMu can achieve favorable performance in EFGs.

CCS Concepts

• **Computing methodologies → Reinforcement learning**; **Regularization**.

Keywords

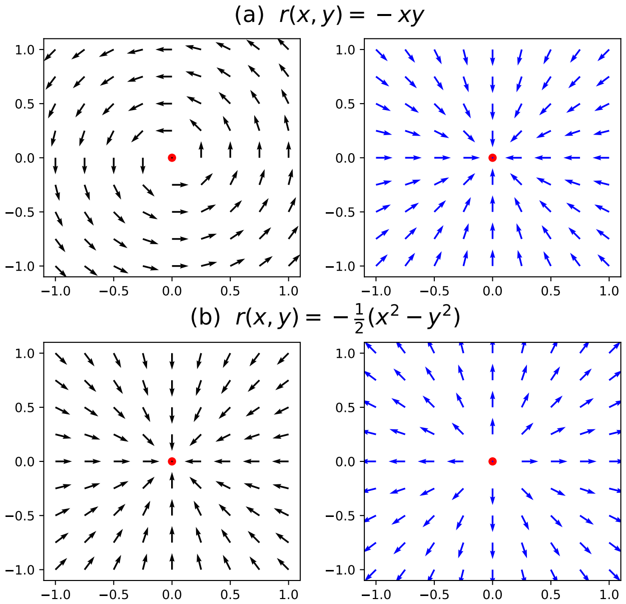
Reinforcement Learning,Regularization,Zero-Sum Games

1 Introduction

Online learning has a rich history that is inextricably intertwined with the advancement of game theory, convex optimization, and machine learning. One of the earliest manifestations of online learning can be attributed to the proposal of fictitious play [10] as a method for solving two-player zero-sum games. Ensuing result [39] has revealed that iteratively computing a best response to each other's history of play in (zero-sum) matrix games leads to convergence to the set of Nash equilibria. This kind of learning paradigm can be linked to the notion of no-regret learning [13], which shares a common historical thread with game theory that dates back to Blackwell's approachability theorem [4, 8].

It is folklore that the time-average policies of no-regret algorithms in self-play converge to a Nash equilibrium in two-player zero-sum games (called *average-iterate convergence*) [13]. A plethora of online learning algorithms, including the celebrated Online Mirror Descent (OMD) [50] and Follow-the-Regularized-Leader (FTRL) [5], ensure that the worst case regret is upper bounded sublinearly with learning iterations [43], thus allowing for a global on-average convergence to the Nash equilibrium over time. A myriad of studies have significantly expanded the applicability of the no-regret theorem to a broader class of settings, covering extensive-form games (EFGs) [24, 55], Markov games (MGs) [6, 46], differential games (i.e. smooth games) [47], and auctions [16].

However, the average-iterate convergence characteristic poses significant challenges in game theory and its practical applications, especially when representing policies using deep networks [22]. In most game settings, averaging neural network weights does not directly correspond to an average of the policies represented by those networks [14, 21]. To mimic agents' average behaviors, it is often required to maintain an additional reservoir buffer, typically hundreds or even thousands of times the size of the game, to store past transition data [22] or historical network parameters [27], leading to extremely high memory demands. Moreover, the average policy cannot be represented precisely due to the inherent neural network approximation error.


(a) $r(x, y) = -xy$

(b) $r(x, y) = -\frac{1}{2}(x^2 - y^2)$

Figure 1: Learning dynamics (black arrows) and second-order dynamics (blue arrows) in the (A) Hamiltonian game, and (B) potential game.

Accordingly, it is imperative to develop no-regret algorithms that converge to (approximate) Nash equilibrium without averaging (called *last-iterate convergence*). However, previous research has demonstrated that the standard no-regret algorithms can lead to cyclic behaviors [36, 54] or even chaotic behaviors [40] of the real-time policy. As shown in the top left side of Figure 1, in a

Hamiltonian game[1], the learning dynamics of the standard no-regret algorithms will cycle around the equilibrium point due to the conflict between the optimization objectives, resulting in convergence failure. In recent literature, methods based on optimistic gradients [15, 51], predictive updates [52], and opponent learning awareness [18] shed lights on how to break the cyclic behaviors by predicting opponents' next moves. Algorithmically, these methods can be considered variations of the optimistic/extra-gradient methods, where the gradient dynamics are modified through the introduction of approximate second-order information [41]. The right plot at the top of Figure 1 shows the convergence direction of the second-order gradient dynamics in the Hamiltonian cycle towards the equilibrium point, which highlights the effectiveness of introducing (approximate) second-order information when jumping out of cycles. However, in different types of games, the second-order information may have different effects on the learning dynamics. As depicted in the bottom plots of Figure 1, in some potential games, the use of second-order information can impede convergence, particularly when the real-time policy exhibits chaotic behavior.

Contributions: The contributions of this paper are mainly three-fold: (1) instead of using the approximate second-order information, we introduce an extra regularization, independent of the game types, into the underlying game to enhance its potential component and thus establish general-case last-iterate convergence; (2) by incorporating regularization, we present a variant of FTRL called **Regularized FTRL** (RegFTRL) that is able to converge at an exponentially fast rate in NFGs without either the optimistic update or the uniqueness assumptions, and investigate two approaches, annealing and adaption approaches, to build algorithms that converge to an exact Nash equilibrium; and (3) we extend RegFTRL to EFGs and propose a model-free reinforcement learning algorithm, **FollowMu**, as a practical implementation of RegFTRL, and validate its performance in Kuhn & Leduc and Phantom Tic-Tac-Toe.

2 Related Work

Research in the realm of last-iterate convergence can be roughly divided into two lines: the optimistic update paradigm and the regularization technique.

In the optimistic update approach, previous studies [14, 30, 34, 38] have investigated the last-iterate convergence in simple unconstrained cases, which are not directly applicable to the NFG/EFG setting. In cases where a unique Nash equilibrium is assumed, Daskalakis and Panageas [15] and Wei et al. [51] extend the scope of research by providing last-iterate convergence guarantees for Optimistic Multiplicative Weights Update (OMWU) (corresponds to Optimistic FTRL with an entropy regularizer) in NFGs, while Wei et al. [51] further prove the convergence of Optimistic Gradient Descent/Ascent (OGDA) (corresponds to optimistic OMD with a L_2 regularizer) without the uniqueness assumption. In the context of EFGs, the pioneering work by Farina et al. [17] empirically demonstrates the last-iterate convergence of OMWU, while Lee et al. [28] subsequently establish theoretical proofs with the uniqueness assumption. Recently, Cai et al. [11], Gorbunov et al. [20] extend the convergence properties of OGDA to monotone games, which include many common classes of games, such as zero-sum polymatrix games and concave-convex games. However, the optimistic update is not flawless. From a theoretical standpoint, OMWU still lacks an explicit last-iterate convergence rate, even in NFGs, without the reliance on uniqueness assumption [28, 51]. Additionally, the analysis of OMWU in EFGs is built over the sequence-form strategy [28, 51], which exhibits limitations in scaling to large games. In practical terms,the implementation of the optimistic update approach often necessitates the computation of multiple strategies at each iteration. Furthermore, OGDA has high per-iteration complexity due to the costly projection operations at each iteration, which adds to the computational burden. In contrast, our proposed approach, RegFTRL, offers distinct advantages. It obviates the requirement for the uniqueness condition and emphasizes the behavior-form strategy, making it more compatible with reinforcement learning and readily adaptable to large-scale games.

Within the realm of learning dynamics, the regularization technique has emerged as pivotal tools for accelerating convergence [1, 2, 37, 44]. Pérolat et al. [37] conduct a comprehensive analysis of the impact of entropy regularization on continuous-time dynamics, and propose a reward transformation method to achieve linear convergence in EFGs using counterfactual values. However, it is imperative to note that their theoretical findings pertaining to continuous-time dynamics do not inherently extend to the desired discrete-time results. Moreover, the use of counterfactual values presents scalability challenges in large-scale settings [32]. Furthermore, their reward transformation technique can lead to estimation issues due to the arbitrarily cumulative sum. Wang et al. [49] show that the GDA algorithm, with a decreasing learning rate, achieves last-iterate convergence in strongly monotone games. In the context of monotone games, the establishment of strong monotonicity is achievable through the incorporation of a strongly convex regularizer. Similar to our work, Magnetic Mirror Descent (MMD) [44] and FTRL with Slingshot Perturbation (FTRL-SP) [1] investigate the influences of general-case regularization on last-iterate convergence, and both provide the linear convergence rate to the regularized equilibrium, albeit FTRL-SP with a more strict restriction on the learning rate. However, MMD exclusively achieves convergence towards an approximated equilibrium in NFGs, while the concurrent work FTRL-SP provides a convergence rate to an exact Nash equilibrium with the L_2 regularizer in monotone games. Furthermore, their analysis could not encompass the behavior-form EFGs considered in our paper. Due to the uniqueness of the quantal response equilibrium (QRE), certain endeavors have attempted to combine optimistic update with additional entropy-regularization to remove the uniqueness assumption associated with OMWU [12, 31], but these approaches still inherit limitations of the optimistic update paradigm.

Compared with the aforementioned related works that either focus on matrix games [2, 44], only consider a special regularization [2], or assume continuous-time feedback [37], we go one step further and prove that RegFTRL converges to an exact Nash equilibrium via general-case regularization in NFGs. Additionally, through empirical observations, we substantiate that RegFTRL equipped with alternative regularization, consistently exhibits last-iterate convergence in behavior-form EFGs.

[1] See Appendix C for discussions on Hamiltonian and potential games. In brief, an NFG is potential if there is a single potential function g such that $V_{\pi^1,\pi^2} - V_{\hat{\pi}^1,\pi^2} = -g(\pi^1,\pi^2) + g(\hat{\pi}^1,\pi^2)$ for all $\pi^1, \hat{\pi}^1, \pi^2$.

3 Preliminaries

3.1 Normal-Form Game

NFGs is representing scenarios where only one stage exists and all the players act simultaneously. Each player i selects action $a^i \in \mathcal{A}$, and then player 1 receives a reward $r(a^1, a^2) \in [0, 1]$ while player 2 receives a reward $-r(a^1, a^2)$. For a given policy $\pi = (\pi^1, \pi^2) \in \prod_{i=1}^{2} \Delta_{\mathcal{A}}$, the Q-function for player 1 is defined as $Q_\pi(a^1) = \mathbb{E}_{a^2 \sim \pi^2}[r(a^1, a^2)]$, and the value function as $V_\pi = \mathbb{E}_{a^1 \sim \pi^1}[Q_\pi(a^1)]$. The value functions of player 2 are the negative values of player 1.

FTRL (See Appendix C for a more detailed introduction) is an intuitive algorithm: for player i, at each time step it maximizes the sum of the past returns $y_t^i = \int_0^t (2 \cdot \mathbf{1}_{1=i} - 1) Q_{\pi_k} dk$ with a regularization $\psi : \Delta_{\mathcal{A}} \to \mathbb{R}$, i.e., $\pi_{t+1}^i = \arg\max_{p \in \Delta_{\mathcal{A}}} [\eta \langle p, y_t^i \rangle - \psi(p)]$ where $\langle \cdot, \cdot \rangle$ means inner product and $\eta > 0$ is the learning rate.

A Nash equilibrium is a widely used solution concept for games. In a Nash equilibrium, no player can improve his/her expected utility by deviating from his/her specified strategy. In two-player zero-sum normal-form games, a strategy profile $\pi^* = (\pi_1^*, \pi_2^*)$ is called a Nash equilibrium if for any$\pi_1 \in \Delta(A_1)$ and $\pi_2 \in \Delta(A_2)$,

$$v_1^{\pi_1^*, \pi_2} \geq v_1^{\pi_1^*, \pi_2^*} \geq v_1^{\pi_1, \pi_2^*}$$

We denote the set of Nash equilibria by Π_*. An ϵ-Nash equilibrium (π_1, π_2) is and approximation of a Nash equilibrium, which satisfies the following inequality:

$$\max_{\widetilde{\pi^1} \in \Delta_{\mathcal{A}}} V_{\widetilde{\pi^1}, \pi^2} + \max_{\widetilde{\pi^2} \in \Delta_{\mathcal{A}}} V_{\pi^1, \widetilde{\pi^2}} \leq \epsilon$$

3.2 Extensive-Form Games

EFGs can be viewed as an extension of NFGs. The representation of a two-player zero-sum EFG $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{P}, \{r^h\}_{h=1}^{H} \rangle$ is based on a game tree of depth H, modeling the sequential interactions involving each player $i \in \mathcal{N} = \{1, 2\}$ and a chance player c. At each history $s \in \mathcal{S}$ at time $h \in [H]$, corresponding to a node at level h in the finite rooted game tree, the player function $\mathcal{P}(s) \mapsto i \in \mathcal{N} \cup \{c\}$ determines a player or chance to play an action $a \in \mathcal{A}(s)$. As a result, player 1 will receive a reward $r^h(s, a) \in [0, 1]$ (and player 2 will receive a reward $-r^h(s, a)$), and the history will transition to its successor history $s' = sa$ at time $h + 1$. We denote $s' \sqsubset s$ if s' is led from s. Due to the imperfect-information, at each history s, only an **information state** $I \in \mathcal{I}$ can be observed, where histories $s \in I$ are indistinguishable for the current player. We use $I(s)$ to denote the information state I corresponding to a history s.

A **behavior-form strategy** $\pi(a|I)$ is defined on each information state: $\pi(\cdot|I) \mapsto \Delta_{\mathcal{A}(I)}$($\Delta$ is a probability simplex, and Δ° means the interior of Δ). We further denote the restriction of π over $\mathcal{I}_i \subseteq \mathcal{I}$ by π^i, and thus $\pi = (\pi^i, \pi^{-i})$. If all players follow π, the **reach probability** of a history s can be computed by $\rho^\pi(s) = \prod_{s'a \sqsubset s} \pi(a|I(s'))$. Thus we have:

$$r^h(I, a) = \sum_{s \in I, a} \rho^\pi(s, a) r^h(s, a) / \sum_{s \in I} \rho^\pi(s), \tag{1}$$

where $\rho^\pi(s, a) = \rho^\pi(s)\pi(a|I(s))$. For player 1, the value function is defined as $V_\pi^h(I) = \mathbb{E}\Big[\sum_{h'=h}^{H} [r^{h'}(I_{h'}, a_{h'})]\Big]$, and Q-function as $Q_\pi^h(I, a) = r^h(I, a) + \mathbb{E}_{I'=I(ha), h \in I}[V_\pi^{h+1}(I')]$. The value functions of player 2 are the negative one of player 1.

3.3 Other Notations

For a strictly convex and continuously differentiable function ψ, we denote the Bregman divergence as $D_\psi(p, q) = \psi(p) - \psi(q) - \langle \nabla \psi(q), p - q \rangle$, and the Kullback-Leibler divergence (i.e., Bregman divergence with $\psi(p) = \sum_a p(a) \ln p(a)$) as D_{KL}. Then, we say that ψ is λ-strongly convex with respect to $\|\cdot\|$ if $D_\psi(p, q) \geq \frac{\lambda}{2} \|p - q\|^2$, and g is λ-strongly convex relative to ψ if $\langle \nabla g(p) - \nabla g(q), p - q \rangle \geq \lambda \langle \nabla \psi(p) - \nabla \psi(q), p - q \rangle$. Note that ψ and $D_\psi(\cdot, q)$ is 1-strongly convex relative to ψ.

4 Stabilize the Learning Dynamics via Regularization

This section utilizes the regularization to stabilize the learning dynamics of FTRL in general games, and presents a comprehensive study of its last-iterate convergence in NFGs. All proofs of our theoretical results are given in Appendix A.

4.1 Last-Iterate Convergence in NFGs

Since the learning dynamics of FTRL will converge in potential games [23] but cycle in Hamiltonian games [7, 33], an intuitive idea to stabilize the FTRL dynamics is to add an extra potential component to the underlying game, i.e., to enhance the potential component of the original game. In this subsection, we present this potential enhancement method in NFGs. With an arbitrary potential function g, we consider the potential-enhancement optimization problem:

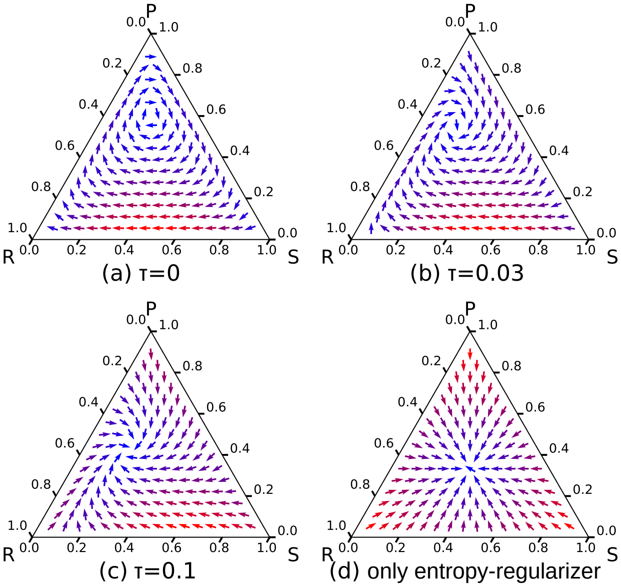
$$\max_{\pi^1 \in \Delta_{\mathcal{A}}} \min_{\pi^2 \in \Delta_{\mathcal{A}}} V_{\pi^1, \pi^2} - \tau g(\pi^1, \pi^2), \tag{2}$$

where $\tau > 0$ is a weight parameter to control the strength of the additional potential component, and thus the original game can be obtained by setting $\tau = 0$. We further consider a decentralized potential function, i.e., $g(\pi^1, \pi^2) = g^1(\pi^1) - g^2(\pi^2)$, in terms of ease-of-use. It can be found that problem (2) is a generalization of entropy-regularized problem by setting $g^i(p) = \langle p, \ln p \rangle$. Inspired by this, we develop **RegFTRL** by incorporating the additional potential function into FTRL:

$$\pi_t^i = \arg\max_{p \in \Delta_A} [\eta \langle p, y_t^i \rangle - \psi(p)], \tag{3}$$

$$y_t^i(a) = \int_0^t \left[\delta^i Q_{\pi_k}(a) - \tau [\nabla g^i(\pi_k^i)]_a\right] dk, \delta^i = 2 \cdot \mathbf{1}_{1=i} - 1,$$

where $\eta > 0$ is the learning rate, and the regularization function $\psi : \Delta_{\mathcal{A}} \to \mathbb{R}$ is strictly convex and continuously differentiable on $\Delta_{\mathcal{A}}$. Note that $\int_0^t \left[\delta^i Q_{\pi_k}(a) - \tau [\nabla g^i(\pi_k^i)]_a\right] dk = \sum_{k=0}^{t-1} \left[\delta^i Q_{\pi_k}(a) - \tau [\nabla g^i(\pi_k^i)]_a\right]$ under discrete-time settings, and FTRL can be obtained by setting $\tau = 0$. It can be found that the learning dynamics of RegFTRL in the original game is equivalent with the FTRL dynamics in the potential-enhancement optimization problem (2). In RegFTRL, the extra potential term is expected to force the dynamics to escape from cycles. Figure 2 shows this insight visually, which describes RegFTRL dynamics in a simple Hamiltonian game of biased Rock-Paper-Scissors. Here we take $g^i(\pi^i) = D_{\mathrm{KL}}(\mu^i, \pi^i)$ (μ is



(a) τ=0 (b) τ=0.03 (c) τ=0.1 (d) only entropy-regularizer

Figure 2: The vector fields of RegFTRL with varying weights in biased Rock-Paper-Scissors.

an uniform strategy) as an example. As can be seen in Figure 2 (a), the FTRL dynamics cycle and fail to converge to the interior Nash equilibrium, while the RegFTRL dynamics with only the potential term (i.e., $y_t^i = -\int_0^t \nabla g^i(\pi_k^i)dk$) directly converge to a stationary point in Figure 2 (d), although this stationary point is independent of the original game. However, with some small weight parameters, RegFTRL flows towards a stationary point near the Nash equilibrium as shown in Figure 2 (b) and (c).

We now analyze the theoretical properties of RegFTRL. Before that, we make some necessary assumptions.

Assumption 1 (Well Defined). *Assume ψ is 1-strongly convex with respect to $\|\cdot\|$ and $\pi_t \in \mathcal{B} = \mathcal{B}^1 \times \mathcal{B}^2 \subseteq \prod_{i=1}^2 \Delta_{\mathcal{A}}^\circ$, where $\{\pi_t\}_{t\geq 0}$ is generated by RegFTRL.*

Assumption 2. *For $i \in \{1, 2\}$, assume g^i is continuously differentiable and λ-strongly convex relative to ψ over $\Delta_{\mathcal{A}}^\circ$. This also implies that ∇g^i is L-smooth over $\mathcal{B}$, i.e., $\|\nabla g^i(p) - \nabla g^i(q)\| \leq L\|p - q\|$ for $\forall p, q \in \mathcal{B}$. Furthermore, we assume g^i has an interior minimum point $\mu^i \in \mathcal{B}^i$, and re-denote g^i as g_μ^i for the sake of clarity. We call this minimum point μ as **reference strategy**.*

Assumption 3 (Regularized Equilibrium). *Assume $\pi_\mu \in \mathcal{B}$ is the interior stationary point of continuous-time RegFTRL dynamics with $\psi(p) = \langle p, \ln p\rangle$ and g_μ.*

Assumption 1 is to ensure that π_t generated by RegFTRL is well defined. This assumption is also required in MMD [44]. Assumption 2 allows RegFTRL to get fast convergence via leveraging the curvature of g_μ, and then the potential enhancement method actually is the regularization technique. Assumption 3 is guaranteed when $g_\mu^i(\pi) = D_{\text{KL}}(\mu^i, \pi^i)$ or $g_\mu^i(\pi) = D_{\text{KL}}(\pi^i, \mu^i)$ (more details refer to Appendix A). Under these assumptions, we first present the properties of the regularized equilibrium, and then give the linear convergence guarantees of RegFTRL without the uniqueness condition in both continuous-time and discrete-time settings.

Theorem 1. *Under Assumption 1~3, the regularized equilibrium $\pi_\mu \in \mathcal{B}$ satisfies: (1) π_μ is unique; and (2) π_μ is an ϵ-Nash equilibrium, where $\epsilon = \mathcal{E}(\pi_\mu) = \tau \sum_{i=1}^2 \left(\max_a [\nabla g^i(\pi_\mu^i)]_a - \langle \pi_\mu^i, \nabla g^i(\pi_\mu^i)\rangle\right) \geq 0$.*

Theorem 2. *Given Assumption 1~3, the **continuous-time** π_t generated by continuous-time version of RegFTRL dynamics satisfies:*

$$D_\psi(\pi_\mu, \pi_t) \leq D_\psi(\pi_\mu, \pi_0) \exp(-\eta\tau\lambda \cdot t),$$

*while the **discrete-time** π_t generated by discrete-time version of RegFTRL dynamics satisfies:*

$$D_\psi(\pi_\mu, \pi_t) \leq D_\psi(\pi_\mu, \pi_0)(1 + \eta\tau\lambda)^{-t}, \text{ and}$$

$$\mathcal{E}(\pi_t) \leq \mathcal{E}(\pi_\mu) + 2\sqrt{D_\psi(\pi_\mu, \pi_0)}(1 + \eta\tau\lambda)^{-t/2},$$

if $\psi(p) = \langle p, \ln p\rangle$ and $0 < \eta \leq \frac{\tau\lambda}{\tilde{L}^2}$, where $\tilde{L} = \max\{\tau L, 1\}$.

Remark 1. *Theorem 1 implies that π_μ is a Nash equilibrium if $\mathcal{E}(\pi_\mu) = 0$, which means $\pi_\mu = \mu$. Additionally, by combining Theorem 1 and Theorem 2, it can be also found that the weight parameter τ introduces a trade-off between the speed of convergence and the bias in the Nash equilibrium. These observations inspire us to develop the two approaches mentioned in Section 4.2 to reach an exact Nash equilibrium.*

Remark 2. *Take the quantal response equilibrium (QRE) as an example, Theorem 2 implies that the discrete-time RegFTRL is guaranteed to find an ϵ-QRE in $O(\frac{1}{\ln(1+\eta\tau)} \ln \frac{1}{\epsilon})$ iterations. Besides, FTRL-SP [1] and entropy-regularized OMWU [12] both require $O(\frac{1}{-\ln(1-\eta\tau/2)} \ln \frac{1}{\epsilon})$ iterations.*

4.2 Convergence to an Exact Nash Equilibrium

We next introduce two approaches to find an exact Nash equilibrium. Drawing insights from Theorem 1 , the distance between the regularized equilibrium π_μ and the set of Nash equilibria Π_* can be effectively controlled through the manipulation of the weight parameter τ. Therefore, we first propose the **annealing approach** that gradually decreases weight parameter τ in order to diminish the bias associated with the equilibrium. Nonetheless, it is worth noting that, as implied by Theorem 2, the speed of convergence might be adversely affected as the weight parameter τ undergoes reduction.

Another one is the **adaption approach**, similar to the direct convergence method [37]. As indicated in Remark 1, a Nash equilibrium can be achieved when $\pi_\mu = \mu$, implying that the regularized equilibrium will exhibit closer proximity to Π_* if the reference strategy μ approximates a Nash equilibrium. Therefore, we set reference strategy μ to π_t every N iterations, with the expectation that π_t will progressively converge to a Nash equilibrium. Indeed, with μ_k denoting the k-th reference strategy and a sufficiency large value for N, π_t converges to π_{μ_k} from Theorem 2. Subsequently, the subsequent reference strategy μ_{k+1} is adjusted to coincide with π_{μ_k}. Intuitively, as k increases, π_{μ_k} coincides with μ_k, consequently driving the reference strategies towards convergence with a Nash equilibrium in the underlying game. This intuition is formally substantiated by the following theorem. It is important to highlight that, unlike the annealing approach, the adaption approach obviates the necessity for a diminishing weight parameter τ, thus preserving a consistent convergence rate.

Theorem 3. *If $g_\mu(\pi) = D_\phi(\pi,\mu)$ and $g_\mu(\pi) = D_{KL}(\mu,\pi)$, then for any interior point μ_0, the sequence of reference strategies $\{\mu_k\}_{k\geq 0}$ converges to a Nash equilibrium of the original game.*

5 FollowMu: A Practical Implementation of RegFTRL

Motivated by the effect of RegFTRL in NFGs, this section generalizes RegFTRL to EFGs, and proposes a novel model-free reinforcement learning algorithm, named FollowMu (**Follow** the reference strategy $\boldsymbol{\mu}$), which combines RegFTRL with function approximation techniques. In EFGs, the RegFTRL policy updated rule can be written as:

$$\pi_t(\cdot|I) = \mathop{\arg\max}_{p\in\Delta_{\mathcal{A}(I)}}[\eta\langle p, y_t^{h,\tau}(I,\cdot)\rangle - \psi(p)], \tag{4}$$

$$y_t^{h,\tau}(I,a) = \sum_{k=0}^{t-1}\Big[\delta(I)Q_k^{h,\tau}(I,a) - \tau[\nabla g_\mu^I(\pi_k)]_a\Big],$$

where $\delta(I) = 2\cdot\mathbf{1}_{1=\mathcal{P}(\mathrm{I})} - 1$, $g_\mu^I(\pi_k) := g_\mu(\pi_k(\cdot|I))$, and the value functions are updated as follows:

$$\begin{cases} Q_0 = 0, V_0(I) = -\delta(I)\tau\langle\pi_0(\cdot|I), \nabla g_\mu^I(\pi_0)\rangle \\ Q_t^{h,\tau}(I,a) = r^h(I,a) + \mathbb{E}_{I'=I(ha),h\in I}[V_{t-1}^{h+1,\tau}(I')] \\ V_t^{h,\tau}(I) = (1-\alpha_t)V_{t-1}^{h,\tau}(I) \\ +\alpha_t\sum_a \pi_t(a|I)\big[Q_t^{h,\tau}(I,a) - \delta(I)\tau\cdot\nabla[g_\mu^I(\pi_t)]_a\big]. \end{cases} \tag{5}$$

It can be found that RegFTRL is compatible with the actor-critic framework, wherein the actor is responsible for policy updates through the utilization of RegFTRL (as shown in Eq. (3)), while the critic undertakes the task of value function updates on a relatively slower timescale.

$$\begin{aligned} &\pi_{t+1}(a|I) \propto \exp(z_t(I,a)), \\ &z_t(I,a) \simeq A(I,a;\theta_t) - V(I;\omega_t). \end{aligned} \tag{6}$$

Let $A(I,a;\theta_t)$ be the actor network parameterized by θ_t, and $V(I;\omega_t)$ be the critic network parameterized by ω_t. At time step t, the critic network $V(I;\omega_t)$ is trained to approximate the value function $V_{\pi_t}(I)$ of the real-time strategy, and the actor network $A(I,a;\theta_t)$ is trained to fit the cumulative advantage function of past iterations plus the Q-function of the current-iterate strategy (with the regularized term):

$$\begin{aligned} &A(I,a;\theta_t) \\ &\simeq \sum_{k=0}^{t-1}\left[Q_{\pi_k}(I,a) - V_{\pi_k}(I) - \tau\log\frac{\pi_k(a|I)}{\mu(a|I)}\right] \\ &\quad + Q_{\pi_t}(I,a) - \tau\log\frac{\pi_t(a|I)}{\mu(a|I)} \\ &\simeq [A(I,a;\theta_{t-1}) - V(I;\omega_{t-1})] + G - \tau\log\frac{\pi_t(a|I)}{\mu(a|I)}, \end{aligned} \tag{7}$$

where G is the empirical estimator of $Q_{\pi_t}(I,a)$. Then, if we take ψ to be the entropy regularizer, the next-iterate strategy can be computed by: Here we employ the advantage function $Q_{\pi_t}(I,a) - V_{\pi_t}(I)$, as a substitution for the Q-function $Q_{\pi_t}(I,a)$, for the sake of enhancing numerical stability and robustness. Despite this alteration, the strategy update formulation in Eq.(6) remains equivalent to the updated strategy employed in RegFTRL, attributed to the shift-invariant nature inherent in the softmax function. Meanwhile, the reference strategy will be updated $\mu \leftarrow \pi_t$ every N iterations.

Algorithm 1 FollowMu

```
Initialize: $\pi_0$ as uniform, $\theta_0$, and $\omega_0$ arbitrarily;
for $t = 0, 1, \ldots$ do
    if $t \mod N = 0$ then
        $\mu \leftarrow \pi_t$
    end if
    Collect replay buffer $\mathcal{B}_t \sim \pi_t$
    for $k = 0, 1, \ldots$ do
        Fetch a mini-batch of samples $\mathcal{D}$ from $\mathcal{B}_t$
        for $(I, a) \in \mathcal{D}$ do
            $G \leftarrow \mathrm{Return}(I, a, \mathcal{D})$
            if $t = 0$ then
                $A_{\mathrm{tmp}} \leftarrow 0$
            else
                $A_{\mathrm{tmp}}$
                $\leftarrow \min\{\ell, \max\{0, A(I,a;\theta_{t-1}) - V(I;\omega_{t-1})\}\}$
            end if
            $A_{\mathrm{target}} \leftarrow A_{\mathrm{tmp}} + G - \tau\log\frac{\pi_t(a|I)}{\mu(a|I)}$
            $\theta_t \leftarrow \mathrm{UpdateActor}(I, a, A_{\mathrm{target}})$
        end for
        for $I \in \mathcal{D}$ do
            $G \leftarrow \mathrm{Return}(I, \mathcal{D})$
            $\omega_t \leftarrow \mathrm{UpdateCritic}(I, G)$
        end for
    end for
    $\pi_{t+1}(a|I) \propto \exp(A(I,a;\theta_t) - V(I;\omega_t))$
end for
```

We summarize our implementation of FollowMu in Algorithm 1, where the return G is estimated by the Monte Carlo method, and the loss of actor and critic network are computed by MSE loss. Note that we use the clipped cumulative advantage function in practice:

$$\begin{aligned} A(I,a;\theta_t) = &\min\Big\{\ell, \max\Big\{0, A(I,a;\theta_{t-1}) - V(I;\omega_{t-1})\Big\}\Big\} \\ &+ G - \tau\log\frac{\pi_t(a|I)}{\mu(a|I)}, \end{aligned}$$

where the clipping operator is employed to ensure the stability of the training process, and $\ell > 0$ controls the strength of clipping. The clipping operator $\min\{\ell, x\}$ serves to effectively limit the magnitude of the cumulative advantage function, thereby preventing it from becoming excessively large and leading to performance collapse. Conversely, when dealing with cumulative values that are too small, we employ the positive clipping operator $\max\{0, x\}$ instead of $\max\{-\ell, x\}$ to truncate these values. In fact, this clipping operation is identical to the one used in CFR+ [45], which is a simple yet highly effective technique for improving performance [9].

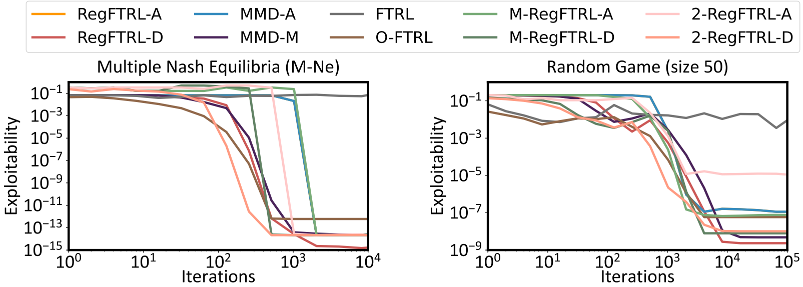


Figure 3: M-NE & Random Game: Exploitability vs. Iterations

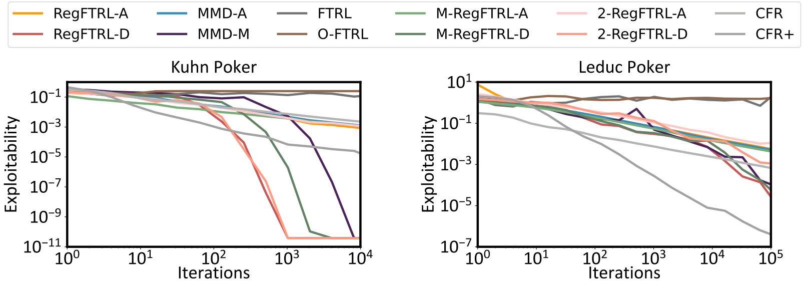


Figure 4: Kuhn & Leduc Poker: Exploitability vs. Iterations

Additionally, when collecting the buffer, the current policy will be perturbed by a small ϵ probability.

6 Experiments

In this section, we validate our methods on NFGs and EFGs utilizing the exploitability metric (i.e., $\mathcal{E}(\pi)$) under two experimental settings, i.e., the full-information feedback setting and the neural-based sample setting. In the full-information feedback setting, we evaluate the performance of RegFTRL as a Nash equilibrium solver, employing the annealing approach and adaption approach. Note that the potential function g_μ in RegFTRL is set to $g_\mu(\pi) = D_{\mathrm{KL}}(\pi, \mu)$. In addition to this, we also consider moment projection $g_\mu(\pi) = D_{\mathrm{KL}}(\mu, \pi)$ and L_2 norm $g_\mu(\pi) = \frac{1}{2}\|\pi - \mu\|_2^2$ for examining the impact brought by different regularization. We abbreviate RegFTRL equipped with $g_\mu(\pi) = D_{\mathrm{KL}}(\mu, \pi)$ as M-RegFTRL, and RegFTRL with $g_\mu(\pi) = \frac{1}{2}\|\pi - \mu\|_2^2$ as 2-RegFTRL. In neural-based sample settings, we assess the efficacy of FollowMu as a deep multi-agent reinforcement learning algorithm through self-play. Further details about experimental settings are included in Appendix B.

6.1 Full-Information Feedback Setting

In this case, we compare the performances of RegFTRL-A (abbr. RegFTRL with annealing approach), RegFTRL-D (abbr. RegFTRL with adaption approach) with baselines: (i) FTRL, (ii) O-FTRL (abbr. optimistic FTRL), (iii) CFR [55], (iv)MMD-A (abbr. MMD with annealing weight), and (v) MMD-M (abbr. MMD with moving magnet). For NFGs, we focus on two games: Multiple Nash Equilibria (abbr. M-Ne) and a random utility game with 50 actions. M-Ne, as introduced in prior work [51], is characterized by a set of Nash equilibria. For the random utility game, the 50×50 payoff matrix is drawn from a standard Gaussian distribution in an i.i.d. manner. For EFGs, we consider games implemented in OpenSpiel [26]: Kuhn Poker and Leduc Poker, with 54 and 9300 non-terminal histories, respectively.

Figure 3 presents the NFG results. Across all games considered, it is evident that FTRL fails to converge to an equilibrium. Conversely, all other algorithms consistently demonstrate linear convergence rates, aligning with theoretical guarantees. This observation underscores the significant impact of the optimistic update paradigm and regularization techniques in facilitating last-iterate convergence. It is noteworthy that RegFTRL shares mathematical equivalence with MMD within the NFG context. Consequently, RegFTRL-A exhibits performance comparable to that of MMD-A. However, an interesting contrast emerges between RegFTRL-D and MMD-M, with the former displaying superior performance. This discrepancy can potentially be attributed to the reference strategy updated by the moving magnet approach, which retains past-iterate strategies, consequently causing it to deviate from the Nash equilibrium.

Figure 4 provides the results observed within Kuhn & Leduc Poker. Unlike the performances in NFGs, in both Poker games, O-FTRL performs poorly, which might be attributed to the behavior-form based implement. In contrast, despite the absence of theoretical convergence guarantees under EFGs, both M-RegFTRL and 2-RegFTRL exhibit an exponentially fast convergence rate. This outcome underscores their potential utility in EFGs, despite the inherent lack of formal guarantees. Furthermore, an interesting trend emerges wherein the adaption approach consistently outperforms the annealing approach in both NFGs and EFGs. This phenomenon can be elucidated by referring to Theorem 2, which indicates that a decaying weight parameter τ can bring a slower convergence rate, aligning with our empirical observations.

6.2 Neural-Based Sample Setting

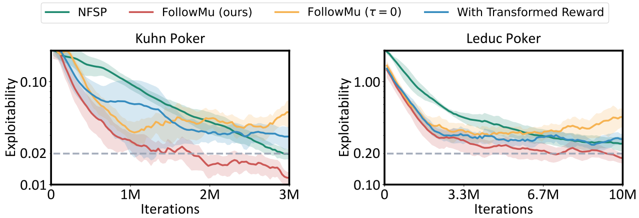


Figure 5: Kuhn & Leduc Poker: Exploitability vs. Iterations

Table 1: Mean±standard deviation of the approximate exploitability on Phantom Tic-Tac-Toe.

	FollowMu	NFSP	PPO	Uniform Agent
1M	0.39 ± 0.02	0.91 ± 0.04	0.94 ± 0.03	0.79 ± 0.02
10M	0.21 ± 0.04	0.80 ± 0.03	0.91 ± 0.03	0.79 ± 0.02

In this case, we validate our practical implement of RegFTRL, i.e., FollowMu, can work effectively with a function approximator. Within both Poker benchmarks, as depicted in Figure 5, we conduct a comparative analysis involving FollowMu, NFSP [22], FollowMu without regularization (i.e., practical implement of FTRL), and FollowMu with a transformed reward [37]. The results reveal that FollowMu consistently outperforms the other baselines in terms of exploitability. The notable disparity between FollowMu and FollowMu ($\tau = 0$) aligns with the learning dynamics associated with RegFTRL and FTRL, substantiating the effectiveness of the incorporated regularization term. Furthermore, our findings indicate that FollowMu surpasses FollowMu with a transformed reward. This distinction may be attributed to the fact that FollowMu introduces the regularization term at the return level, mitigating the cumulative sum effect encountered at the reward level.

Table 1 reports the performances of FollowMu with NFSP, PPO [42], and Uniform Agent (employing a uniform strategy consistently) on Phantom Tic-Tac-Toe in OpenSpiel, which is an imperfect-information game where the winner receives a payoff of +1 and the losing player receives −1. The evaluation of approximate exploitability in Phantom Tic-Tac-Toe is computed through a trained DQN best response, owing to the substantial scale of the game. The outcomes underscore that both FollowMu and NFSP exhibit enhanced performance following 10 million steps of training compared to their performance after 1 million steps. In contrast, PPO exhibits negligible improvement, consistent with the fact that it is designed for single-agent environments. Notably, FollowMu stands out as the top performer, significantly outperforming the baselines.

7 Conclusion

In this paper, we introduce RegFTRL to enhance the stability of FTRL dynamics through a general-case regularization, and establish the last-iterate linear convergence in NFGs without either the uniqueness condition or the optimistic update paradigm. Furthermore, our investigation extends to probing the feasibility of achieving convergence towards an exact Nash equilibrium through two straightforward yet highly efficient approaches. Additionally, we extend RegFTRL to EFGs, and propose a model-free reinforcement learning algorithm for zero-sum games, named FollowMu. The numerical simulation reveals that RegFTRL outperforms FTRL and O-FTRL in various zero-sum games, and FollowMu attains favorable performance levels against strong baselines. Future research could investigate its analyses in General-Sum Games and explore its application in more complex scenarios, such as Multiplayer Poker.

Acknowledgments

We gratefully acknowledge support from the National Natural Science Foundation of China (No. 62076259), the Fundamental and Applicational Research Funds of Guangdong province (No. 2023A1515012946), and the Fundamental Research Funds for the Central Universities-Sun Yat-sen University.

References

[1] Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. 2023. A Slingshot Approach to Learning in Monotone Games. *arXiv* abs/2305.16610 (2023).

[2] Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki. 2022. Last-Iterate Convergence with Full- and Noisy-Information Feedback in Two-Player Zero-Sum Games. *arXiv* abs/2208.09855 (2022).

[3] Kenshi Abe, Mitsuki Sakamoto, and Atsushi Iwasaki. 2022. Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *Uncertainty in Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 180)*. 1–10.

[4] Jacob Abernethy, Peter L Bartlett, and Elad Hazan. 2011. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 27–46.

[5] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. 2008. Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization. In *Annual Conference on Learning Theory*. 263–274.

[6] Yu Bai, Chi Jin, and Tiancheng Yu. 2020. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems* 33 (2020), 2159–2170.

[7] David Balduzzi, Sébastien Racanière, James Martens, Jakob N. Foerster, Karl Tuyls, and Thore Graepel. 2018. The Mechanics of n-Player Differentiable Games. In *International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. 363–372.

[8] David Blackwell. 1956. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.* 6 (1956), 1–8.

[9] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. 2017. Heads-up limit hold'em poker is solved. *Commun. ACM* 60, 11 (2017), 81–88.

[10] George W Brown. 1949. *Some notes on computation of games solutions.* Technical Report. RAND CORP SANTA MONICA CA.

[11] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. 2022. Finite-Time Last-Iterate Convergence for Learning in Multi-Player Games. In *Advances in Neural Information Processing Systems*.

[12] Shicong Cen, Yuting Wei, and Yuejie Chi. 2021. Fast Policy Extragradient Methods for Competitive Games with Entropy Regularization. In *Advances in Neural Information Processing Systems*. 27952–27964.

[13] Nicolò Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, learning, and games.* Cambridge University Press.

[14] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. 2018. Training GANs with Optimism. In *International Conference on Learning Representations*.

[15] Constantinos Daskalakis and Ioannis Panageas. 2019. Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization. In *Information Technology Convergence and Services (LIPIcs, Vol. 124)*. 27:1–27:18.

[16] Xiaotie Deng, Xinyan Hu, Tao Lin, and Weiqiang Zheng. 2022. Nash convergence of mean-based learning algorithms in first price auctions. In *Proceedings of the ACM Web Conference 2022*. 141–150.

[17] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. 2019. Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions. In *Advances in Neural Information Processing Systems*. 5222–5232.

[18] Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with Opponent-Learning Awareness. In *International Conference on Autonomous Agents and MultiAgent Systems*. 122–130.

[19] Daniel Friedman. 1991. Evolutionary games in economics. *Econometrica: journal of the econometric society* (1991), 637–666.

[20] Eduard Gorbunov, Adrien B. Taylor, and Gauthier Gidel. 2022. Last-Iterate Convergence of Optimistic Gradient Method for Monotone Variational Inequalities. In *Advances in Neural Information Processing Systems*.

[21] Johannes Heinrich, Marc Lanctot, and David Silver. 2015. Fictitious self-play in extensive-form games. In *International Conference on Machine Learning*. 805–813.

[22] Johannes Heinrich and David Silver. 2016. Deep Reinforcement Learning from Self-Play in Imperfect-Information Games. *arXiv* abs/1603.01121 (2016).

[23] Amélie Héliou, Johanne Cohen, and Panayotis Mertikopoulos. 2017. Learning with Bandit Feedback in Potential Games. In *Advances in Neural Information Processing Systems*. 6369–6378.

[24] Samid Hoda, Andrew Gilpin, Javier F. Pena, and Tuomas Sandholm. 2010. Smoothing Techniques for Computing Nash Equilibria of Sequential Games. *Mathematics of Operations Research* 35 (2010), 494–512.

[25] Josef Hofbauer and Karl Sigmund. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press.

[26] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinícius Flores Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas W. Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. 2019. OpenSpiel: A Framework for Reinforcement Learning in Games. *arXiv* abs/1908.09453 (2019).

[27] Marc Lanctot, Vinícius Flores Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*. 4190–4203.

[28] Chung-Wei Lee, Christian Kroer, and Haipeng Luo. 2021. Last-iterate Convergence in Extensive-Form Games. In *Advances in Neural Information Processing Systems*. 14293–14305.

[29] Alistair Letcher, David Balduzzi, Sébastien Racanière, James Martens, Jakob N. Foerster, Karl Tuyls, and Thore Graepel. 2019. Differentiable Game Mechanics. *The Journal of Machine Learning Research* 20 (2019), 3032–3071.

[30] Tengyuan Liang and James Stokes. 2019. Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks. In *International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*. 907–915.

[31] Mingyang Liu, Asuman E. Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. 2022. The Power of Regularization in Solving Extensive-Form Games. *arXiv* abs/2206.09495 (2022).

[32] Stephen Marcus McAleer, Gabriele Farina, Marc Lanctot, and Tuomas Sandholm. 2023. ESCHER: Eschewing Importance Sampling in Games by Computing a History Value Function to Estimate Regret. In *International Conference on Learning Representations*.

[33] Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. 2018. Cycles in Adversarial Regularized Learning. In *ACM-SIAM Symposium on Discrete Algorithms*. 2703–2717.

[34] Aryan Mokhtari, Asuman E. Ozdaglar, and Sarath Pattathil. 2020. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. In *International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*. 1497–1507.

[35] Dov Monderer and Lloyd S Shapley. 1996. Potential games. *Games and Economic Behavior* 14, 1 (1996), 124–143.

[36] Shayegan Omidshafiei, Christos H. Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Pérolat, and Rémi Munos. 2019. α-Rank: Multi-Agent Evaluation by Evolution. *Scientific reports* 9, 1 (2019), 9937.

[37] Julien Pérolat, Rémi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro A. Ortega, Neil Burch, Thomas W. Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. 2021. From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization. In *International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. 8525–8535.

[38] Alexander Rakhlin and Karthik Sridharan. 2013. Optimization, Learning, and Games with Predictable Sequences. In *Advances in Neural Information Processing System*, Vol. 26.

[39] Julia Robinson. 1951. An iterative method of solving a game. *Annals of Mathematics* (1951), 296–301.

[40] Yuzuru Sato, Eizo Akiyama, and J Doyne Farmer. 2002. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences* 99, 7 (2002), 4748–4751.

[41] Florian Schäfer and Anima Anandkumar. 2019. Competitive Gradient Descent. In *Advances in Neural Information Processing Systems*. 7623–7633.

[42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv* abs/1707.06347 (2017).

[43] Shai Shalev-Shwartz. 2012. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning* 4 (2012), 107–194.

[44] Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. 2023. A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games. In *International Conference on Learning Representations*.

[45] Oskari Tammelin. 2014. Solving Large Imperfect Information Games Using CFR+. *arXiv* abs/1407.5042 (2014).

[46] Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. 2021. Online Learning in Unknown Markov Games. In *International Conference on Machine Learning*. 10279–10288.

[47] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. 2019. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. *Advances in Neural Information Processing Systems* 32 (2019).

[48] Michael Walton and Viliam Lisý. 2021. Multi-agent Reinforcement Learning in OpenSpiel: A Reproduction Report. *arXiv* abs/2103.00187 (2021).

[49] Zifan Wang, Yi Shen, Michael Zavlanos, and Karl Henrik Johansson. 2022. No-Regret Learning in Strongly Monotone Games Converges to a Nash Equilibrium. (2022).

[50] Manfred K Warmuth, Arun K Jagota, et al. 1997. Continuous and discrete-time nonlinear gradient descent: Relative loss bounds and convergence. In *International Symposium on Artificial Intelligence and Mathematics*, Vol. 326. Citeseer.

[51] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. 2021. Linear Last-iterate Convergence in Constrained Saddle-point Optimization. In *International Conference on Learning Representations*.

[52] Abhay Kumar Yadav, Sohil Shah, Zheng Xu, David W. Jacobs, and Tom Goldstein. 2018. Stabilizing Adversarial Nets with Prediction Methods. In *International Conference on Learning Representations*.

[53] E Christopher Zeeman. 2006. Population dynamics from game theory. In *Global Theory of Dynamical Systems: Proceedings of an International Conference Held at Northwestern University*. Springer, 471–497.

[54] Martin Zinkevich, Amy Greenwald, and Michael L. Littman. 2005. Cyclic Equilibria in Markov Games. In *Advances in Neural Information Processing Systems*. 1641–1648.

[55] Martin Zinkevich, Michael Johanson, Michael H. Bowling, and Carmelo Piccione. 2007. Regret Minimization in Games with Incomplete Information. In *Advances in Neural Information Processing Systems*. 1729–1736.

A Theoretical guarantee

This section will provide relevant theoretical supplements for the theorems presented in section 4. For convenience, we denote $Q^1_\pi(a) = Q_\pi(a)$ and $Q^2_\pi(a) = -Q_\pi(a)$.

Proposition 1. *Assumption 3 is guaranteed when $g_\mu(\pi) = D_{KL}(\pi,\mu)$ and $g_\mu(\pi) = D_{KL}(\mu,\pi)$.*

Proof. For any interior reference strategy μ, if there exists an action a_0 such that $\pi_\mu(a_0) = 0$, then $Q_{\pi_\mu}(a_0) - \tau[\nabla g_\mu(\pi_\mu)]_{a_0} < Q_{\pi_\mu}(a_*) - \tau[\nabla g_\mu(\pi_\mu)]_{a_*}$ holds for any $a_* \in \{a \in \mathcal{A} | \pi_\mu(a) > 0\}$. However, for $g_\mu(\pi) = D_{\mathrm{KL}}(\pi,\mu)$ and $g_\mu(\pi) = D_{\mathrm{KL}}(\mu,\pi)$, we have:

$$
\begin{aligned}
& Q_{\pi_\mu}(a_0) - \tau[\nabla g_\mu(\pi_\mu)]_{a_0} \\
=& \begin{cases} Q_{\pi_\mu}(a_0) - \tau(\ln\frac{\pi_\mu(a_0)}{\mu(a_0)} + 1) = \infty, & \text{if } g_\mu(\pi) = D_{\mathrm{KL}}(\pi,\mu), \\ Q_{\pi_\mu}(a_0) + \tau\frac{\mu(a_0)}{\pi_\mu(a_0)} = \infty, & \text{if } g_\mu(\pi) = D_{\mathrm{KL}}(\mu,\pi), \end{cases}
\end{aligned}
$$

which is a contradiction, and thus Assumption 3 holds. □

We continue by proving the properties of regularized equilibrium, and the last-iterate convergence of continuous-time version of RegFTRL in NFGs.

A.1 Proof of Theorem 1

Proof. Note that Theorem 2 holds for all regularized equilibria. This means that π_μ is unique if the weight parameter τ and the reference strategy μ are fixed. Thus we only need to provide the proofs of the second statement that π_μ is an ϵ-Nash equilibrium.

Since regularized equilibrium is the interior stationary point of continuous-time RegFTRL dynamics with $\psi(p) = \langle p, \ln p\rangle$. By the method of Lagrange multiplier, it can be found that the dynamics defined by continuous-time RegFTRL with the entropy regularizer is equivalent to the following dynamics:

$$
\frac{d}{dt}\pi^i_t(a) = \pi^i_t(a)\big(Q^i_{\pi_t}(a) - \tau[\nabla g^i_\mu(\pi^i_t)]_a - V^i_{\pi_t} + \tau\langle \pi^i_t, \nabla g^i_\mu(\pi^i_t)\rangle\big) \tag{8}
$$

From Assumption 3, we have:

$$
Q^i_{\pi_\mu}(a) - \tau[\nabla g^i_\mu(\pi^i_\mu)]_a - V^i_{\pi_\mu} + \tau\langle \pi^i_\mu, \nabla g^i_\mu(\pi^i_\mu)\rangle = 0 \tag{9}
$$

Therefore, we have:

$$
\begin{aligned}
\mathcal{E}(\pi_\mu) = \sum_{i=1}^{2} \max_{p^i\in\Delta_\mathcal{A}} V^i_{p,\pi^{-i}_\mu} &= \sum_{i=1}^{2}\big[\max_{p^i\in\Delta_\mathcal{A}} V^i_{p,\pi^{-i}_\mu} - V^i_{\pi_\mu}\big] \\
&= \sum_{i=1}^{2}\big[\max_{a\in\mathcal{A}} Q^i_{\pi_\mu}(a) - V^i_{\pi_\mu}\big] \\
&\le \tau\sum_{i=1}^{2}\big[\max_{a\in\mathcal{A}}[\nabla g^i_\mu(\pi^i_\mu)]_a - \langle \pi^i_\mu, \nabla g^i_\mu(\pi^i_\mu)\rangle\big]
\end{aligned}
$$

Thus, the proof is completed. □

A.2 Proof of Theorem 2

Proof. Theorem 2 is the summary of lemma 1, Theorem 4 and Theorem 5, and thus we omit the proofs here. □

Lemma 1. *$\mathcal{E}$ can be bounded as follows:*

$$
\mathcal{E}(\pi_t) \le \mathcal{E}(\pi_\mu) + 2\sqrt{D_{KL}(\pi_\mu,\pi_t)}.
$$

Proof. From the definition of $\mathcal{E}$, we have:

$$
\begin{aligned}
\mathcal{E}(\pi_t) &= \sum_{i=1}^{2} \max_{p^i\in\Delta_\mathcal{A}} V^i_{p,\pi^{-i}_t} \\
&= \sum_{i=1}^{2} \big(\max_{p^i\in\Delta_\mathcal{A}} V^i_{p,\pi^{-i}_\mu} + \max_{p^i\in\Delta_\mathcal{A}} V^i_{p,\pi^{-i}_t} - \max_{p^i\in\Delta_\mathcal{A}} V^i_{p,\pi^{-i}_\mu}\big) \\
&= \mathcal{E}(\pi_\mu) + \sum_{i=1}^{2} \big(\max_{p^i\in\Delta_\mathcal{A}} V^i_{p,\pi^{-i}_t} - \max_{p^i\in\Delta_\mathcal{A}} V^i_{p,\pi^{-i}_\mu}\big) \\
&\le \mathcal{E}(\pi_\mu) + \sum_{i=1}^{2} \max_{p^i\in\Delta_\mathcal{A}} \big(V^i_{p,\pi^{-i}_t} - V^i_{p,\pi^{-i}_\mu}\big) \\
&\overset{\text{Hölder's inequality}}{\le} \mathcal{E}(\pi_\mu) + \sum_{i=1}^{2} \big(\|\pi^i_\mu - \pi^i_t\|_1 \max_{p^{-i}\in\Delta_\mathcal{A}} \|Q^i_{\pi^i_t,p^{-i}}\|_\infty\big) \\
&\overset{\text{Pinsker inequality}}{\le} \mathcal{E}(\pi_\mu) + \sum_{i=1}^{2} \sqrt{2D_{\mathrm{KL}}(\pi^i_\mu,\pi^i_t)} \\
&\overset{\text{Cauchy inequality}}{\le} \mathcal{E}(\pi_\mu) + \sqrt{2}\sqrt{2\sum_{i=1}^{2} D_{\mathrm{KL}}(\pi^i_\mu,\pi^i_t)} \\
&= \mathcal{E}(\pi_\mu) + 2\sqrt{D_{\mathrm{KL}}(\pi_\mu,\pi_t)}.
\end{aligned}
$$

□

A.2.1 *Continuous-Time RegFTRL.*

Theorem 4. *Let Assumption 1~3 hold. Then, π_t generated by continuous-time version of RegFTRL dynamics satisfies:*

$$
D_\psi(\pi_\mu,\pi_t) \le D_\psi(\pi_\mu,\pi_0)\cdot \exp(-\eta\tau\lambda\cdot t).
$$

Proof. By lemma C.1 in [3], we have:

$$
\begin{aligned}
\frac{d}{dt}D_\psi(\pi_\mu,\pi_t) &= \sum_i \Big\langle \frac{d}{dt}y^i_t, \pi^i_t - \pi^i_\mu\Big\rangle \\
&= \eta\sum_i \Big\langle Q^i_{\pi_t} - \tau\nabla g^i_\mu(\pi^i_t), \pi^i_t - \pi^i_\mu\Big\rangle \\
&= \eta\sum_i \Big\{ V^i_{\pi_t} - V^i_{\pi^i_\mu,\pi^{-i}_t} - \tau\langle \nabla g^i_\mu(\pi^i_t), \pi^i_t - \pi^i_\mu\rangle\Big\} \\
&= \eta\sum_i \Big\{ V^i_{\pi^i_t,\pi^{-i}_\mu} - \tau\langle \nabla g^i_\mu(\pi^i_t), \pi^i_t - \pi^i_\mu\rangle\Big\} \\
&= \eta\sum_i \Big\{ \langle \pi^i_t, Q^i_{\pi_\mu}\rangle - \tau\langle \nabla g^i_\mu(\pi^i_t), \pi^i_t - \pi^i_\mu\rangle\Big\} \\
&= \eta\sum_i \Big\{ V^i_{\pi_\mu} - \tau\langle \nabla g^i_\mu(\pi^i_\mu), \pi^i_\mu\rangle \\
&\qquad + \tau\langle \nabla g^i_\mu(\pi^i_\mu), \pi^i_t\rangle - \tau\langle \nabla g^i_\mu(\pi^i_t), \pi^i_t - \pi^i_\mu\rangle\Big\} \\
&= -\eta\tau\sum_i \Big\{ \langle \nabla g^i_\mu(\pi^i_\mu) - \nabla g^i_\mu(\pi^i_t), \pi^i_\mu - \pi^i_t\Big\rangle \\
&\le -\eta\tau\lambda\sum_i \Big\{ \langle \nabla\psi(\pi^i_\mu) - \nabla\psi(\pi^i_t), \pi^i_\mu - \pi^i_t\Big\rangle \\
&= -\eta\tau\lambda\Big[D_\psi(\pi_\mu,\pi_t) + D_\psi(\pi_t,\pi_\mu)\Big] \\
&\le -\eta\tau\lambda D_\psi(\pi_\mu,\pi_t).
\end{aligned} \tag{10}
$$

The sixth equality follows from Eq.(9). Therefore, we have:

$$D_{\psi}(\pi_\mu, \pi_t) \leq D_{\psi}(\pi_\mu, \pi_0) \cdot \exp(-\eta\tau\lambda \cdot t).$$

Thus, the proof is completed. □

A.2.2 Discrete-Time RegFTRL.

THEOREM 5. *Let Assumption 1∼3 hold. Then, π_t generated by discrete-time version of RegFTRL dynamics satisfies:*

$$D_{\psi}(\pi_\mu, \pi_t) \leq D_{\psi}(\pi_\mu, \pi_0) \cdot (1 + \eta\tau\lambda)^{-t},$$

if $\psi(p) = \langle p, \ln p\rangle$ and $0 < \eta \leq \frac{\tau\lambda}{\tilde{L}^2}$, where $\tilde{L} = \max\{\tau L, 1\}$.

PROOF. Let us define $f^i_\pi := Q^i_\pi - \tau \nabla g^i_\mu(\pi^i)$. With $\psi(p) = \langle p, \ln p\rangle$ and the method of Lagrange multiplier, π_t generated by RegFTRL satisfies:

$$\begin{aligned}
& \pi^i_{t+1}(a) \propto \pi^i_t(a) \exp\left\{\eta\left[Q^i_{\pi_t}(a) - \tau[\nabla g^i_\mu(\pi^i_t)]_a\right]\right\} \\
\iff \quad & \pi^i_{t+1} = \operatorname*{arg\,max}_{p \in \Delta_{\mathcal{A}}} \left\{\eta\langle p, f^i_{\pi_t}\rangle - D_\psi(p, \pi^i_t)\right\} \\
\iff \quad & \left\langle \eta f^i_{\pi_t} - \nabla\psi(\pi^i_{t+1}) + \nabla\psi(\pi^i_t), \pi^i - \pi^i_{t+1}\right\rangle \leq 0, \quad \forall \pi^i \in \Delta_{\mathcal{A}} \\
\iff \quad & \left\langle \eta f^i_{\pi_t}, \pi^i - \pi^i_{t+1}\right\rangle \leq \left\langle \nabla\psi(\pi^i_{t+1}) - \nabla\psi(\pi^i_t), \pi^i - \pi^i_{t+1}\right\rangle, \\
& \hfill \forall \pi^i \in \Delta_{\mathcal{A}} \\
\iff \quad & \left\langle \eta f^i_{\pi_t}, \pi^i - \pi^i_{t+1}\right\rangle \\
& \leq D_{\mathrm{KL}}(\pi^i, \pi^i_t) - D_{\mathrm{KL}}(\pi^i, \pi^i_{t+1}) - D_{\mathrm{KL}}(\pi^i_{t+1}, \pi^i_t), \\
& \hfill \forall \pi^i \in \Delta_{\mathcal{A}}
\end{aligned}$$

The third "$\iff$" follows from the equivalent first order optimality conditions. Therefore, we have:

$$D_{\mathrm{KL}}(\pi^i_\mu, \pi^i_{t+1}) \leq D_{\mathrm{KL}}(\pi^i_\mu, \pi^i_t) - D_{\mathrm{KL}}(\pi^i_{t+1}, \pi^i_t) - \eta\left\langle f^i_{\pi_t}, \pi^i_\mu - \pi^i_{t+1}\right\rangle. \tag{11}$$

On the other hand, we have:

$$\begin{aligned}
& \eta \sum_{i=1}^{2} \left\langle f^i_{\pi_t}, \pi^i_\mu - \pi^i_t\right\rangle \\
=& \eta \sum_{i=1}^{2} \left\langle f^i_{\pi_\mu}, \pi^i_\mu - \pi^i_t\right\rangle + \eta \sum_{i=1}^{2} \left\langle Q^i_{\pi_t} - Q^i_{\pi_\mu}, \pi^i_\mu - \pi^i_t\right\rangle \\
& + \eta\tau \sum_{i=1}^{2} \left\langle \nabla g^i_\mu(\pi^i_\mu) - \nabla g^i_\mu(\pi^i_t), \pi^i_\mu - \pi^i_t\right\rangle \\
\geq& \eta\tau\lambda \sum_{i=1}^{2} \left\langle \nabla\psi(\pi^i_\mu) - \nabla\psi(\pi^i_t), \pi^i_\mu - \pi^i_t\right\rangle \\
=& \eta\tau\lambda\left[D_{\mathrm{KL}}(\pi_\mu, \pi_t) + D_{\mathrm{KL}}(\pi_t, \pi_\mu)\right].
\end{aligned} \tag{12}$$

The inequality follows from the fact that

$$\left\langle f^i_{\pi_\mu}, \pi^i_\mu - \pi^i_t\right\rangle = \left\langle V^i_{\pi_\mu}\mathbf{1} - \tau\langle \pi^i_\mu, \nabla g^i_\mu(\pi^i_\mu)\rangle\mathbf{1}, \pi^i_\mu - \pi^i_t\right\rangle = 0,$$

and

$$\sum_{i=1}^{2} \left\langle Q^i_{\pi_t} - Q^i_{\pi_\mu}, \pi^i_\mu - \pi^i_t\right\rangle = \sum_{i=1}^{2} (V^i_{\pi^i_\mu, \pi^{-i}_t} - V^i_{\pi_\mu} - V^i_{\pi_t} + V^i_{\pi^i_t, \pi^{-i}_\mu}) = 0.$$

By combining Eq.(11) and Eq.(12), we have:

$$\begin{aligned}
& D_{\mathrm{KL}}(\pi_\mu, \pi_{t+1}) \\
\leq& D_{\mathrm{KL}}(\pi_\mu, \pi_t) - D_{\mathrm{KL}}(\pi_{t+1}, \pi_t) \\
& - \eta \sum_{i=1}^{2} \left\langle f^i_{\pi_t}, \pi^i_\mu - \pi^i_{t+1}\right\rangle \\
\leq& D_{\mathrm{KL}}(\pi_\mu, \pi_t) - D_{\mathrm{KL}}(\pi_{t+1}, \pi_t) - \eta\tau\lambda D_{\mathrm{KL}}(\pi_\mu, \pi_{t+1}) \\
& - \eta\tau\lambda D_{\mathrm{KL}}(\pi_{t+1}, \pi_\mu) + \eta \sum_{i=1}^{2} \left\langle f^i_{\pi_{t+1}} - f^i_{\pi_t}, \pi^i_\mu - \pi^i_{t+1}\right\rangle \\
\leq& D_{\mathrm{KL}}(\pi_\mu, \pi_t) - D_{\mathrm{KL}}(\pi_{t+1}, \pi_t) \\
& - \eta\tau\lambda D_{\mathrm{KL}}(\pi_\mu, \pi_{t+1}) - \eta\tau\lambda D_{\mathrm{KL}}(\pi_{t+1}, \pi_\mu) \\
& + \eta \max\{1, \tau L\}\|\pi_{t+1} - \pi_t\|_1 \cdot \|\pi_{t+1} - \pi_\mu\|_1 \\
\leq& D_{\mathrm{KL}}(\pi_\mu, \pi_t) - D_{\mathrm{KL}}(\pi_{t+1}, \pi_t) \\
& - \eta\tau\lambda D_{\mathrm{KL}}(\pi_\mu, \pi_{t+1}) - \eta\tau\lambda D_{\mathrm{KL}}(\pi_{t+1}, \pi_\mu) \\
& + \frac{1}{2}\|\pi_{t+1} - \pi_t\|_1^2 + \frac{\eta^2\tilde{L}^2}{2}\|\pi_{t+1} - \pi_\mu\|_1^2 \\
\leq& D_{\mathrm{KL}}(\pi_\mu, \pi_t) - D_{\mathrm{KL}}(\pi_{t+1}, \pi_t) \\
& - \eta\tau\lambda D_{\mathrm{KL}}(\pi_\mu, \pi_{t+1}) - \eta\tau\lambda D_{\mathrm{KL}}(\pi_{t+1}, \pi_\mu) \\
& + D_{\mathrm{KL}}(\pi_{t+1}, \pi_t) + \eta^2\tilde{L}^2 D_{\mathrm{KL}}(\pi_{t+1}, \pi_\mu) \\
\leq& D_{\mathrm{KL}}(\pi_\mu, \pi_t) - \eta\tau\lambda D_{\mathrm{KL}}(\pi_\mu, \pi_{t+1}).
\end{aligned} \tag{13}$$

The third inequality follows from the fact that $r(a^1, a^2) \in [0, 1]$ and ∇g is L-smooth. Next, we successively apply the inequalities $\tilde{L} = \max 1, \tau L$, $2ab \leq \rho a^2 + \frac{b^2}{\rho}$, Pinsker's inequality, and $\eta \leq \frac{\tau\lambda}{\tilde{L}^2}$. Therefore, we have:

$$D_{\mathrm{KL}}(\pi_\mu, \pi_{t+1}) \leq \frac{1}{1 + \eta\tau\lambda} D_{\mathrm{KL}}(\pi_\mu, \pi_t). \tag{14}$$

Thus, the proof is completed. □

A.3 Proof of Theorem 3

We begin with following useful lemmas.

LEMMA 2. *If $\pi_\mu \neq \mu$, we have $D_\phi(\pi_*, \pi_\mu) < D_\phi(\pi_*, \mu), \forall \pi_* \in \Pi_*$.*

PROOF. From the definition of the regularized equilibrium, we have

$$\begin{aligned}
& D_\phi(\pi_*, \pi_\mu) - D_\phi(\pi_*, \mu) \\
=& - D_\phi(\pi_\mu, \mu) - \sum_{i=1}^{2} \langle \nabla\phi(\pi^i_\mu) - \nabla\phi(\mu^i), \pi^i_* - \pi^i_\mu\rangle \\
=& - D_\phi(\pi_\mu, \mu) - \frac{1}{\tau}\sum_{i=1}^{2} \langle Q^i_{\pi_\mu}, \pi^i_* - \pi^i_\mu\rangle \\
\leq& - D_\phi(\pi_\mu, \mu) < 0.
\end{aligned}$$

The inequality follows from the fact that

$$-\sum_{i=1}^{2} \langle Q^i_{\pi_\mu}, \pi^i_* - \pi^i_\mu\rangle = \sum_{i=1}^{2} [V^i_{\pi^i_\mu, \pi^{-i}_*} - V^i_{\pi_*}] \leq 0.$$

Thus, the proof is completed. □

LEMMA 3. *If $\pi_\mu = \mu$, then μ is a Nash equilibrium of the original game.*

Proof. By the definition of the regularized equilibrium, when $\pi_\mu = \mu$, we have:

$$\pi^i_\mu(a)\big[Q^i_{\pi_\mu}(a) - \tau[\nabla g^i_\mu(\pi^i_\mu)]_a - V^i_{\pi_\mu} + \tau\langle \pi^i_\mu, \nabla g^i_\mu(\pi^i_\mu)\rangle\big] = 0$$
$$\overset{\pi_\mu(a)>0}{\Longrightarrow} \quad Q^i_{\pi_\mu}(a) - \tau[\nabla g^i_\mu(\pi^i_\mu)]_a - V^i_{\pi_\mu} + \tau\langle \pi^i_\mu, \nabla g^i_\mu(\pi^i_\mu)\rangle = 0$$
$$\overset{\pi_\mu=\mu}{\Longrightarrow} \quad Q^i_{\pi_\mu}(a) - V^i_{\pi_\mu} = 0$$

Therefore, $V^i_{\pi_\mu} = \max_{a\Delta_{\mathcal{A}}} Q^i_{\pi_\mu}(a)$ for $i = 1, 2$, which means that each player's strategy is a best response to the strategy of the other player. Thus, μ is a Nash equilibrium of the original game. □

Lemma 4. *For any $k \geq 0$, if $\mu_k \in \prod_{i=1}^2 \Delta^\circ_{\mathcal{A}} \backslash \Pi_*$, then $\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_{k+1}) < \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_k)$. Otherwise, if $\mu_k \in \Pi_*$, then $\mu_{k+1} = \mu_k \in \Pi_*$.*

Proof. From lemma 3, if $\mu \in \prod_{i=1}^2 \Delta^\circ_{\mathcal{A}} \backslash \Pi_*$, then we have $\pi_\mu \neq \mu$. Denote $\pi_* = \arg\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu)$. Then from lemma 2, if $\mu \neq \pi_\mu$, we have:

$$\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu) = D_\phi(\pi_*, \mu) > D_\phi(\pi_*, \pi_\mu) \geq \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \pi_\mu).$$

Therefore, we prove the first statement of the lemma. Then we assume that $\mu \in \Pi_*$ implies $\pi_\mu \neq \mu$. From lemma 2, we have $D_\phi(\pi_*, \mu) > D_\phi(\pi_*, \pi_\mu)$ for any $\pi_* \in \Pi_*$, and thus $0 > D_\phi(\mu, \pi_\mu)$ due to $r \in \Pi_*$. It is a contradiction.

Thus, the proof is completed. □

Lemma 5. *Let $F(\mu) = \pi_\mu$ be a map that maps the reference strategy μ to its corresponding regularized equilibrium π_μ. Then, F is continuous.*

Proof. For any given reference strategies $\mu, \hat{\mu} \in \prod_i \Delta^\circ_{\mathcal{A}}$, we denote their associated stationary points as $\pi_\mu, \pi_{\hat{\mu}}$ respectively. Suppose that π_t is the updated strategy of continues-time FTRL dynamics with reference strategy μ and $\psi(p) = \langle p, \ln p\rangle$, then

$$\frac{d}{dt}\pi^i_t(a) = \pi^i_t(a)\big[Q^i_{\pi_t}(a) - \tau[\nabla g^i_\mu(\pi^i_t)]_a - V^i_{\pi_t} + \tau\langle \pi^i_t, \nabla g^i_\mu(\pi^i_t)\rangle\big].$$

Therefore, we have

$$\begin{aligned}\frac{d}{dt}D_{\mathrm{KL}}(\pi_{\hat{\mu}}, \pi_t) &= -\sum_{i=1}^2 \langle \pi^i_{\hat{\mu}}, \frac{1}{\pi^i_t}\frac{d}{dt}\pi^i_t\rangle \\ &= \sum_{i=1}^2 \langle \pi^i_t - \pi^i_{\hat{\mu}}, Q^i_{\pi_t} - \tau\nabla g^i_\mu(\pi^i_t)\rangle \\ &= \underbrace{\sum_{i=1}^2 \langle \pi^i_t - \pi^i_{\hat{\mu}}, Q^i_{\pi_t} - \tau[\nabla\phi(\pi^i_t) - \nabla\phi(\hat{\mu}^i)]\rangle}_{(1)} \\ &\quad + \underbrace{\tau\sum_{i=1}^2 \langle \pi^i_t - \pi^i_{\hat{\mu}}, \nabla\phi(\mu^i) - \nabla\phi(\hat{\mu}^i)\rangle}_{(2)}.\end{aligned}$$

Then,

$$\begin{aligned}(1) &= \sum_{i=1}^2 V_{\pi^i_t, \pi^{-i}_{\hat{\mu}}} - \tau\sum_{i=1}^2 \langle \pi^i_t - \pi^i_{\hat{\mu}}, \nabla\phi(\pi^i_t) - \nabla\phi(\hat{\mu}^i)\rangle \\ &= \sum_{i=1}^2 \langle \pi^i_t, Q^i_{\pi_{\hat{\mu}}}\rangle - \tau\sum_{i=1}^2 \langle \pi^i_t - \pi^i_{\hat{\mu}}, \nabla\phi(\pi^i_t) - \nabla\phi(\hat{\mu}^i)\rangle \\ &= \sum_{i=1}^2 \langle \pi^i_t, V^i_{\pi_{\hat{\mu}}}\rangle + \tau\sum_{i=1}^2 \langle \pi^i_t - \pi^i_{\hat{\mu}}, \nabla\phi(\pi^i_{\hat{\mu}}) - \nabla\phi(\hat{\mu}^i)\rangle \\ &\quad - \tau\sum_{i=1}^2 \langle \pi^i_t - \pi^i_{\hat{\mu}}, \nabla\phi(\pi^i_t) - \nabla\phi(\hat{\mu}^i)\rangle \\ &= \tau\sum_{i=1}^2 \langle \pi^i_t - \pi^i_{\hat{\mu}}, \nabla\phi(\pi^i_{\hat{\mu}}) - \nabla\phi(\pi^i_t)\rangle \\ &= -\tau D_\phi(\pi_t, \pi_{\hat{\mu}}) - \tau D_\phi(\pi_{\hat{\mu}}, \pi_t).\end{aligned}$$

On the other hand, we have $(2) \leq 2\tau L\|\hat{\mu} - \mu\|$. By setting $\pi_t = \pi_\mu$, we have $\pi_t = \pi_\mu$, for any $t \geq 0$, and thus we have $D_\phi(\pi_\mu, \pi_{\hat{\mu}}) \leq 2L\|\hat{\mu} - \mu\|$, which means that F is continuous.

Thus, the proof is completed. □

Proof of Theorem 3

Proof. In the case that $g_\mu(\pi) = D_{\mathrm{KL}}(\mu, \pi)$, RegFTRL is equivalent with M2WU, and thus the convergence result can be guaranteed by Theorem 6.1 in [2]. We next provide the proof of the case that $g_\mu(\pi) = D_\phi(\pi, \mu)$. Denote $b = \lim_{k\to\infty} \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_k) \geq 0$. We next prove that $b = 0$ and thus μ_k converges to Π_*.

By contradiction, we suppose that $b > 0$ and define $B = \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_0)$. From lemma 4, $\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_k)$ monotonically decreases, and thus each μ_k falls into the set $\Omega_{b,B} = \{\mu \in \prod_{i=1}^2 \Delta^\circ_{\mathcal{A}} : b \leq \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu) \leq B\}$. From lemma 5, $\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu)$ is continuous on $\prod_{i=1}^2 \Delta^\circ_{\mathcal{A}(I)}$, and thus $\Omega_{b,B}$ is a compact set due to the boundedness of $\prod_{i=1}^2 \Delta^\circ_{\mathcal{A}}$.

From lemma 5, $\Delta V(\mu) := \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, F(\mu)) - \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu)$ is also continuous. Thus $\Delta V(\mu)$ has a maximum over a compact set, i.e., $M = \max_{\mu\in\Omega_{b,B}} \Delta V(\mu)$ exists. From Lemma 4, $M < 0$, and thus we have:

$$\begin{aligned}\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_k) &= \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_0) \\ &\quad + \sum_{l=0}^{k-1}\Big(\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_{l+1}) - \min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_l)\Big) \\ &\leq B + kM.\end{aligned}$$

This implies that $\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_k) < 0$ for $k > \frac{-B}{M}$, which is a contradiction since $\min_{\pi_*\in\Pi_*} D_\phi(\pi_*, \mu_k) \geq 0$.

Thus, the proof is completed. □

Table 2: Hyper-Parameter Settings of RegFTRL in M-NE/Random Game.

	learning rate η	regularization parameter τ	update period N
RegFTRL-A	$\frac{10}{1+10\times\tau}, \frac{10}{1+10\times\tau}$	$10^{1-t/500}, 10^{1-t/300}$	$0, 0$
RegFTRL-D	$\frac{15}{1+15\times\tau}, \frac{6}{1+6\times\tau}$	$\frac{15}{t}, \frac{6}{t}$	$10, 100$
M-RegFTRL-A	$10, 10$	$10^{1-t/500}, 10^{1-t/300}$	$0, 0$
M-RegFTRL-D	$10, 6$	$\frac{10}{t}, \frac{6}{t}$	$10, 100$
2-RegFTRL-A	$10, 20$	$10^{1-t/500}, 10^{1-t/300}$	$0, 0$
2-RegFTRL-D	$10, 6.5$	$\frac{10}{t}, \frac{6}{t}$	$10, 100$

Table 3: Hyper-Parameter Settings of RegFTRL in Kuhn/Leduc Poker.

	learning rate η	regularization parameter τ	update period N
RegFTRL-A	$\frac{1}{\sqrt{t}}, \frac{1}{\sqrt{t}}$	$\frac{1}{\sqrt{t}}, \frac{5}{\sqrt{t}}$	$0, 0$
RegFTRL-D	$0.3, 0.11$	$0.1, 1$	$30, 30$
M-RegFTRL-A	$\frac{1}{\sqrt{t}}, \frac{1}{\sqrt{t}}$	$\frac{0.5}{\sqrt{t}}, \frac{3}{\sqrt{t}}$	$0, 0$
M-RegFTRL-D	$0.1, 0.11$	$0.1, 1$	$30, 30$
2-RegFTRL-A	$\frac{1}{\sqrt{t}}, \frac{1}{\sqrt{t}}$	$\frac{2}{\sqrt{t}}, \frac{20}{\sqrt{t}}$	$0, 0$
2-RegFTRL-D	$0.1, 0.11$	$0.5, 7$	$30, 30$

B Experimental Settings

B.1 Full-Information Feedback Setting

The payoff matrices of M-NE from [51] is as follows:

	y_1	y_2	y_3	y_4	y_5
x_1	0	1	-1	0	0
x_2	-1	0	1	0	0
x_3	1	-1	0	0	0
x_4	1	-1	0	-2	1
x_5	1	-1	0	1	-2

M-NE has the following set of Nash equilibria:
$\Pi^1_* = \{(1/3, 1/3, 1/3, 0, 0)\}, \Pi^2_* = \{y \in \Delta^5 | y_1 = y_2 = y_3; y_5/2 \leq y_4 \leq 2y_5\}$. For the random utility game, the 50×50 payoff matrix is drawn from a standard Gaussian distribution in an i.i.d. manner. The benchmarks of Kuhn/Leduc Poker is from OpenSpiel. The hyper-parameters for RegFTRL in NFGs are listed in Table 2, and the hyper-parameters in EFGs are listed in Table 3.

B.2 Neural-Based Sample Setting

The benchmarks of Kuhn/Leduc Poker and the implementation of NFSP are all from OpenSpiel, and all the experiments are run on A30. The hyper-parameters for FollowMu are listed in Table 4, while those for NFSP are listed in Table 5, which are referenced from the report [48].

Table 4: Hyper-Parameter Settings of FollowMu in Kuhn/Leduc Poker.

Parameter	Value
hidden_layers_sizes	[128, 128]
batch_size	1024
mini_batch_size	128/256
logit_learning_rate	0.001/0.0005
critic_learning_rate	0.005
max_global_gradient_norm	10.0
optimizer_str	sgd
eta	0.2
refer_policy_update_every	200/500
clip_strength	100

C Additional Preliminaries

C.1 Game Decomposition

Several recent works have shown that an arbitrary game (normal-form type or differential-form type) can be uniquely decomposed into a sum of Hamiltonian and potential components through the generalized Helmholtz decomposition theorem [7, 29]. There are thus two "pure" games: Hamiltonian games (only the Hamiltonian component is present) and potential games (only the potential component). Hamiltonian games, such as Rock-Paper-Scissors, are actually divergence-free vector fields where the cyclic behaviors arise [7]. Hence, FTRL will get stuck in cycles around equilibrium if the Hamiltonian component of the underlying game is dominant. On the other hand, a game is a potential game if there is a single potential function g such that $V_{\pi^1,\pi^2} - V_{\hat{\pi}^1,\pi^2} = -g(\pi^1, \pi^2) + g(\hat{\pi}^1, \pi^2)$ for all $\pi^1, \hat{\pi}^1, \pi^2$. Potential games are well-studied because they can be solved by following the gradient dynamics [7, 35].

C.2 Follow-the-Regularized-Leader

FTRL is an intuitive algorithm: at each time step it maximizes the sum of the past returns with a regularization. For conciseness, we only present the definition of FTRL in NFGs here. Formally, FTRL

Table 5: Hyper-Parameter Settings of NFSP in Kuhn/Leduc Poker.

Parameter	Value
hidden_layers_sizes	[128, 128]
replay_buffer_capacity	200000
reservoir_buffer_capacity	2000000
min_buffer_size_to_learn	1000
anticipatory_param	1
batch_size	128
learn_every	128
rl_learning_rate	0.01
sl_learning_rate	0.01
optimizer_str	sgd
update_target_network_every	19200
discount_factor	1.0
epsilon_decay_duration	10000000
epsilon_start	0.06
epsilon_end	0.001

dynamics is defined as follows:

$$\pi_t^i = \arg\max_{p\in\Delta_{\mathcal{A}}}[\eta\langle p, y_t^i\rangle - \psi_i(p)], \tag{15}$$

$$y_t^i(a) = \int_0^t \delta^i \cdot Q_{\pi_k}(a)dk, \quad \delta^i = 2\cdot \mathbf{1}_{i=1} - 1,$$

where $\langle\cdot,\cdot\rangle$ means inner product, $\eta > 0$ is the learning rate, and the regularization function $\psi : \Delta_{\mathcal{A}} \to \mathbb{R}$ is strictly convex and continuously differentiable on $\Delta_{\mathcal{A}}$. Note that $\int_0^t Q_{\pi_k}(a)dk = \sum_{k=0}^{t-1} Q_{\pi_k}(a)$ under discrete-time settings.

Two prototypical examples of FTRL can be yielded by choosing different regularizers: 1) Replicator Dynamics (RD) induced by the entropy regularizer $\psi_i(p) = \sum_a p(a)\ln p(a)$; and 2) Projection Dynamics (PD) induced by the (square) Euclidean regularizer $\psi_i(p) = \frac{1}{2}\sum_a |p(a)|^2$ [33].

RD is an important learning dynamics studied in evolution game theory [25, 53], where the central focus is to mimic the population's evolution process. The dynamics of RD can be given by the following differential equation:

$$\frac{d}{dt}\pi_t^i(a) = \pi_t^i(a)\delta^i(Q_{\pi_t}(a) - V_{\pi_t}). \tag{16}$$

PD is introduced as a geometric model of the evolution of play in population games [19]. Denoting the support set of policy as $\mathrm{supp}(\pi_t^i) = \{a \in \mathcal{A} : \pi_t^i(a) > 0\}$, the dynamics of PD can be defined as follows:

$$\frac{d}{dt}\pi_t^i(a) = \delta^i Q_{\pi_t}(a) - |\mathrm{supp}(\pi_t^i)|^{-1} \sum_{a'\in\mathrm{supp}(\pi_t^i)} \delta^i Q_{\pi_t}(a'), \tag{17}$$

if $a \in \mathrm{supp}(\pi_t^i)$, and $\frac{d}{dt}\pi_t^i(a) = 0$ otherwise.