From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization

Julien Perolat [1] Remi Munos [1] Jean-Baptiste Lespiau [1] Shayegan Omidshafiei [1] Mark Rowland [1]
Pedro Ortega [1] Neil Burch [1] Thomas Anthony [1] David Balduzzi [1] Bart De Vylder [1] Georgios Piliouras [2]
Marc Lanctot [1] Karl Tuyls [1]

Abstract

In this paper we investigate the Follow the Regularized Leader dynamics in sequential imperfect information games (IIG). We generalize existing results of Poincaré recurrence from normal-form games to zero-sum two-player imperfect information games and other sequential game settings. We then investigate how adapting the reward (by adding a regularization term) of the game can give strong convergence guarantees in monotone games. We continue by showing how this reward adaptation technique can be leveraged to build algorithms that converge exactly to the Nash equilibrium. Finally, we show how these insights can be directly used to build state-of-the-art model-free algorithms for zero-sum two-player Imperfect Information Games (IIG).

1. Introduction

This paper addresses the problem of learning a Nash equilibrium in several classes of games. Learning Nash equilibria in competitive games is complex as agents no longer share information but behave independently. Various techniques have been proposed to solve these games, with the current state-of-the-art usually guaranteeing average-time convergence of the learned policy to a Nash equilibrium, but not necessarily convergence of the policy itself to Nash. Unfortunately, these convergence guarantees are not conducive to learning in large games, which rely on general function approximation techniques (e.g., deep neural networks) that are inherently difficult to time-average. Moreover, the real-time behaviors of the policy can be quite distinctive from its time-average counterpart, and can even diverge away from Nash equilibria (Bailey & Piliouras, 2018).

In some adversarial games, Follow the Regularized Leader (FoReL) is known to be convergent if the equilibrium is deterministic, and recurrent if the equilibrium is mixed with full support (Mertikopoulos et al., 2018). A special case of FoReL dynamics is replicator dynamics (Taylor & Jonker, 1978), the main dynamic of evolutionary game theory, whose recurrent behavior in zero-sum games and generalizations is well studied (Piliouras & Shamma, 2014; Boone & Piliouras, 2019). More generally, value-based methods have been well-studied in multi-agent reinforcement learning (Littman, 1994; 2001; Hu & Wellman, 2003) but numerous issues of convergence have been noticed. But a notable empirical finding shows that regularization of Q-learning in matrix games can induce the policy to converge in real-time to a Nash equilibrium (Tuyls et al., 2003; Kaisers & Tuyls, 2010; 2011) or in the replicator dynamics to Quantal-Response-Equilibrium (Ortega & Legg, 2018; McKelvey & Palfrey, 1995; Tuyls & Nowé, 2005). Other theoretical investigations show that softmax best response can guarantee convergence in Q-learning (Leslie & Collins, 2005).

Motivated by these findings, this paper formally analyzes the impact of regularization on learning dynamics, extending beyond the simple case of matrix games and focusing particularly on the application of FoReL to imperfect information games. The contributions of the paper are as follows:

- We generalize the Poincaré recurrence result (Mertikopoulos et al., 2018) to the case of sequential imperfect information games. This proves that strategies can cycle in IIG when using FoReL (e.g., similar to the normal form game case Fig. 1, (a)).
- We prove that changing the reward structure of the game improves convergence guarantees at the cost of slightly modifying the equilibrium of the game (e.g., as in Fig. 1, (b), (c), (d)).
- We show that this reward adaptation method can be used to build a sequence of closer and closer pseudo-solutions converging onto a Nash equilibrium.

[1]DeepMind [2]SUTD. Correspondence to: Julien Perolat <perolat@google.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

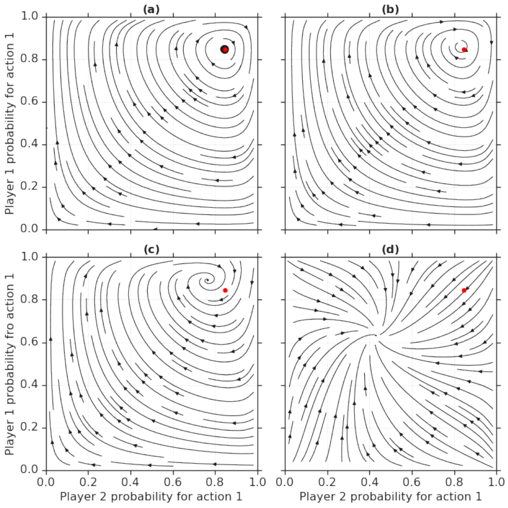

Figure 1. The trajectory plots for FoReL (plot (a)) and for the version with a reward transform with a parameter η multipliers 0.5, 1 and 10 (plots (b), (c), (d) respectively) in a biased matching pennies game (the payoff table for the first player is $[[1,-1],[-1,10]]$). The red dot is the equilibrium policies of the original game.

- We illustrate that by using these theoretical findings, we improve the state-of-the-art of deep reinforcement learning in some imperfect information games.

1.1. Related Work

We discuss related work along three axes: (i) follow the regularized leader and regret minimization, (ii) gradient based methods in differentiable games, and (iii) dynamic programming and reinforcement learning approaches in games.

FoReL and regret minimization in Games. There exists a large corpus of literature providing evidence that minimax equilibria (or Nash equilibria) in **normal form zero-sum two-player** games are often unstable rest points of FoReL (or at best neutrally stable). In evolutionary game theory, the replicator dynamics (Zeeman, 1980; 1981; Weibull, 1997; Gintis, 2009) are known to be unstable in the case of an interior equilibrium in zero-sum two-player normal form games (Bloembergen et al., 2015). Many machine learning approaches can be used in self-play to learn an equilibrium: regret minimization methods have been extensively studied in zero-sum games (Cesa-Biachi & Lugosi, 2006; Syrgkanis et al., 2015; Fudenberg & Levine, 1998; Zinkevich et al., 2008; Hofbauer et al., 2009; Cesa-Biachi & Lugosi, 2006), for which the average policy played over time converges to an equilibrium, but the actual policy is known to be recurrent (Piliouras & Shamma, 2014; Mertikopoulos et al., 2018). The convergence of the actual policy can be obtained when the opponent plays a best response (Waugh & Bagnell, 2015; Abernethy et al., 2018) but not in the self-play setting. The best response sequences of Fictitious Play (Brown, 1951) and smoother variants (Hofbauer & Sandholm, 2002) converge in time-average. **Polymatrix games** can be solved by linear programming (Cai et al., 2016) (we will study a generalization of this class). Regret minimization techniques can be used to learn a Nash equilibrium, [1] but also in this setting, the convergence to a Nash equilibrium requires to compute a time-average policy, and the policy itself is recurrent (Mertikopoulos et al., 2018).

Gradient Based Methods in Differentiable Games. Differentiable games (e.g. GANs) trained by gradient descent present many failure modes (Balduzzi et al., 2018). In (Balduzzi et al., 2018) the authors prove that learning dynamics of gradient descent can cycle in some classes of games. This problem can be resolved by introducing second-order optimization (Balduzzi et al., 2018; Foerster et al., 2017; Mescheder et al., 2017; Letcher et al., 2019), negative momentum (Gidel et al., 2019) or game theoretic algorithms (Oliehoek et al., 2017; Grnarova et al., 2018).

Reinforcement Learning in Games. In sequential imperfect information games RL methods have been applied with mild success. Independent reinforcement learning has many failure modes under these sequential imperfect information settings, as demonstrated in (Lanctot et al., 2017). In zero-sum sequential imperfect information games, the policy can cycle around the minimax equilibrium without ever converging, even in simple single-state games (Piliouras & Shamma, 2014; Mertikopoulos et al., 2018; Singh et al., 2000; Bloembergen et al., 2015; Bailey & Piliouras, 2018). In cooperative settings, players tend to overfit to the opponent while learning, without being able to generalize to other opponents' behaviors (Matignon et al., 2012). Generally speaking, in the sequential setting, learning in games can be addressed by either approximate dynamic programming in the perfect information case (Lagoudakis & Parr, 2002; Pérolat et al., 2015; 2016; Pérolat et al., 2016; 2017; Geist et al., 2019), regret minimization algorithms (Zinkevich et al., 2008; Lanctot, 2013; Lanctot et al., 2009) (which suffer from the aforementioned time-averaging problem), best response algorithms (Heinrich et al., 2015; Lanctot et al., 2017; Heinrich & Silver, 2016), model free reinforcement learning methods (Srinivasan et al., 2018; Heinrich & Silver, 2016) or policy gradient in the worst case (Lockhart et al., 2019). However, the previous model free RL methods are not flawless: Neural Fictitious Self Play (NFSP) (Heinrich & Silver, 2016) maintains two data sets of respectively 600 and 2000 times the size of the game, the methods presented in (Srinivasan et al., 2018) empirically show a convergence in time-average without formal proof, and (Lockhart et al., 2019) require the exact computation of a best response.

[1] Since for a coarse correlated equilibrium, the marginals with respect to the players are a Nash equilibrium (Cai et al., 2016)

2. Warming up: Normal Form Games

We first sketch our main results in repeated zero-sum two-player normal form games.

Background. In a zero-sum two-player normal form game, two players select their actions $a^i \in A$ ($a = (a^1, a^2) = (a^i, a^{-i})$) according to a policy $\pi^i \in \Delta A$ ($\pi = (\pi^1, \pi^2) = (\pi^i, \pi^{-i})$, where $-i$ encodes the opponent of player i), and as a result will receive a reward $r^i_\pi(a^1, a^2)$. The reward is policy-independent ($r^i(a^1, a^2)$) if it is only a function of the actions of the players and not of their policies; policy-independent reward is a standard assumption in the literature. If policy π is played we define the Q-function to be the expected reward for player i for action a_i (i.e. $Q^i_\pi(a^i) = \mathbb{E}_{a^{-i}\sim\pi^{-i}}[r^i_\pi(a^i, a^{-i})]$) and the value function to be the expected reward (i.e. $V^i_\pi = \mathbb{E}_{a\sim\pi}[r^i_\pi(a)] = \mathbb{E}_{a^i\sim\pi^i}[Q^i_\pi(a^i)]$).

By definition, a policy π^* is a **Nash equilibrium** if for all π and for all i we have $V^i_{\pi^i,\pi^{*-i}} - V^i_{\pi^*} \leq 0$. In other words, a Nash equilibrium is a joint policy such that no player has an incentive to change its policy if all the other players stick to their policy.

Follow the Regularized Leader (FoReL). FoReL is an exploration-exploitation algorithm that maximizes the cumulative payoff of the player (exploitation) minus a regularization term (exploration). The continuous time version of this algorithm is defined as follows:

$$y^i_t(a^i) = \int_0^t Q^i_{\pi_s}(a^i)ds \quad \text{and} \quad \pi^i_t = \arg\max_{p\in\Delta A} \Lambda^i(p, y^i_t)$$

where $\Lambda^i(p, y) = \langle y, p\rangle - \phi_i(p)$ and ϕ_i is the regularizer, a function which is assumed to be: (1) continuous and strictly convex on ΔA and (2) smooth on the relative interior of every face of ΔA (including ΔA itself). Standard choices of ϕ_i include: (1) entropy $\phi_i(p) = \sum_a p(a)\log p(a)$, and (2) ℓ_2-norm $\phi_i(p) = \sum_a |p(a)|^2$. The choice of the regularizer lead to different dynamics: entropy regularization yields the replicator dynamics and l_2-norm regularization yields the projection dynamics (Mertikopoulos et al., 2018).

We will write $\phi^*_i(y) = \max_p \Lambda^i(p, y)$ and we have the property that $\arg\max_{p\in\Delta A} \Lambda^i(p, y) = \nabla_y \phi^*_i(y)$ (maximizing argument Shalev-Shwartz et al. (2012, p.147)).

If π^* is a Nash equilibrium, a useful measure of interest that measures the distance to a Nash equilibrium is

$$J(y) = \sum_{i=1}^{2} \left[\phi^*_i(y_i) - \langle y_i, \pi^*_i\rangle\right].$$

This quantity (and its generalization introduced in section. 3) will be used to construct strong Lyapunov functions in many games of interest. As a warm up, this section will explore these convergence results in the normal form case.

Recurrence. If the reward is policy-independent and if there exists an interior equilibrium, it is known that the policy under FoReL will be recurrent (Mertikopoulos et al., 2018). We will generalize this result to sequential Imperfect Information Games in section 4. Crucially, this strong negative result indicates that convergence *cannot* be achieved with FoReL in games with a mixed strategy equilibrium, so long as the reward is policy-independent. Thus, in the rest of the paper, we will explore how to transform the reward by adding a policy-dependent term to guarantee convergence (see section 5,6).

Reward transformation and convergence in normal-form games. Section 5 explores the idea of reward transformation by adding a policy dependent term. If the reward is not policy-independent, one can show that:

$$\frac{d}{dt}J(y) = \sum_{i=1}^{2} \underbrace{[V^i_{\pi^i_t,\pi^{*-i}} - V^i_{\pi^*}]}_{\leq 0 \text{ because } \pi^* \text{ is a Nash}}$$
$$+ \sum_{i=1}^{2} \mathbb{E}_{a\sim(\pi^{*i},\pi^{-i}_t)}[r^i_{\pi^{*i},\pi^{-i}_t}(a) - r^i_{\pi_t}(a)]$$

We later generalize this result in lemma 3.1. As an example, consider the following policy dependent reward, which also preserves the zero-sum property for any policy μ with a full support:

$$r^i_\pi(a) = r^i(a^i, a^{-i}) - \eta\log\frac{\pi^i(a^i)}{\mu^i(a^i)} + \eta\log\frac{\pi^{-i}(a^{-i})}{\mu^{-i}(a^{-i})}$$

Given the above reward, we can show that:

$$\frac{d}{dt}J(y) = \sum_{i=1}^{2} \underbrace{[V^i_{\pi^i_t,\pi^{*-i}} - V^i_{\pi^*}]}_{\leq 0 \text{ because } \pi^* \text{ is a Nash}} - \eta\sum_{i=1}^{2} KL(\pi^{*i}, \pi^i_t)$$

This inequality ensures that π_t will converge to π^*, the Nash of the game defined by $r^i_\pi(a)$, using Lyapunov arguments. Note that π^* will depend on μ and η. Transforming the reward improves the convergence property of the game but will shift the equilibrium, a phenomena illustrated in figure 1 where ϕ_i is the entropy for all players. Thus, this technique does not directly guarantee convergence to the Nash of the original game. We next introduce a technique to adapt the policy-dependent term in the reward, thereby guaranteeing convergence to the actual Nash equilibrium of the game.

Direct Convergence. Solving the original game can be achieved by iteratively solving the game with the reward $r^i_{k,\pi}(h, a) = r^i(a^i, a^{-i}) - \eta\log\frac{\pi^i(a^i)}{\pi^i_{k-1}(a^i)} + \eta\log\frac{\pi^{-i}(a^{-i})}{\pi^{-i}_{k-1}(a^{-i})}$

and use the Nash of that game π_k to modify the reward of the next game (starting with π_0 as the uniform policy). The sequence of policies $(\pi_k)_{k\geq 0}$ converges to π^*, the equilibrium of the policy-independent reward $r^i(a^i, a^{-i})$. Specifically, we can show that:

$$\sum_{i=1}^{2}\Big[KL(\pi^{*i},\pi^i_k) - KL(\pi^{*i},\pi^i_{k-1})\Big] \leq -\sum_{i=1}^{2} KL(\pi^i_k,\pi^i_{k-1})$$

which is enough to prove that $(\pi_k)_{k\geq 0}$ converges to π^*, using Lyapunov-style arguments (this result is proved in section 6). This set of results establishes a foundation for convergent learning in the normal-form case. We next lay out the principles necessary for generalizing to the IIG setting, with our main result detailed in section 6.

3. Background in Sequential Imperfect Information Games

In a sequential imperfect information game, N players and a chance player (written c) interact sequentially starting from a history h_{init}. The set of all possible histories is written $H = \cup_{i\in\{1,\ldots,N,c\}} H_i$. The sets H_i are the set of histories at player's i turn (all H_i are disjoint). The set of terminal histories $\mathcal{Z}_i$ is a subset of H_i in which the game has ended ($\mathcal{Z} = \cup_{i\in\{1,\ldots,N,c\}}\mathcal{Z}_i$). In each history $h \in H$, the current player will observe an information state $x \in \mathcal{X} = \cup_{i\in\{1,\ldots,N,c\}}\mathcal{X}_i$. The function $\tau(h) \mapsto i \in \{1,\ldots,N,c\}$ provides the player's turn at a given history. We will also write $x(h) \in \mathcal{X}$ for the information state corresponding to an history h. We will write $h \in x$ if $x(h) = x$.

At each history $h \in H\backslash\mathcal{Z}$, the current player will play an action $a \in A$. As a result, each player $i \in \{1,\ldots,N\}$ will receive a reward $r^i(h,a)$ and the state will transition to $h' = ha$. We will write $h \sqsubset h'$ if there exists a sequence of k actions $(a_i)_{0\leq i\leq k}$ such that $ha_0\cdots a_k = h'$. The history h is then said to be a prefix of h'.

A policy $\pi(a|x)$ maps an information state x to a distribution over actions ΔA. The restriction of π over $\mathcal{X}_i$ is written π^i and π^{-i} is the restriction of π over $\mathcal{X}\backslash\mathcal{X}_i$. We will write $\pi = (\pi^i, \pi^{-i})$. As in section. 2, we consider a **policy dependent reward** (written $r^i_\pi(h,a)$), which can be dependent on the full policy. The rest of this section introduces reinforcement learning tools used to define FoReL in IIG and used in the proofs.

Value function on the histories. The value of a policy for player i at history h is defined as follow:

$$V^i_\pi(h) = \mathbb{E}\big[\sum_{n\geq 0} r^i_\pi(h_n,a_n)|h_0 = h,\ h_{n+1} = h_n\ a_n,$$
$$a_n \sim \pi(.|x(h_n))\big] = \sum_a \pi(a|x(h))\left[r^i_\pi(h,a) + V^i_\pi(ha)\right]$$

The value of a policy for player i at history h while taking action a is defined as follow:

$$Q^i_\pi(h,a) = \mathbb{E}\big[\sum_{n\geq 0} r^i_\pi(h_n,a_n)|h_0 = h,\ a_0 = a,$$
$$h_{n+1} = h_n\ a_n,\ a_n \sim \pi(.|x(h_n))\big] = r^i_\pi(h,a) + V^i_\pi(ha)$$

Reach probabilities. The reach probability of a history h is (note that this product may include the chance player):

$$\rho^\pi(h) = \prod_{h'a\sqsubset h} \pi(a|x(h'))$$

The reach probability of player i of a history h is:

$$\rho^{\pi^i}(h) = \prod_{h'a\sqsubset h,\ \tau(h')=i} \pi(a|x(h'))$$

The reach probability of player $-i$ of a history h is (this product may include the chance player too):

$$\rho^{\pi^{-i}}(h) = \prod_{h'a\sqsubset h,\ \tau(h')\neq i} \pi(a|x(h'))$$

In the end, $\forall h \in H$:$\rho^\pi(h) = \rho^{\pi^i}(h)\rho^{\pi^{-i}}(h)$

The reach probability of an information state $x \in \mathcal{X}$ is defined as follows:

$$\rho^\pi(x) = \sum_{h\in x}\rho^\pi(h) \text{ and } \rho^{\pi^{-i}}(x) = \sum_{h\in x}\rho^{\pi^{-i}}(h)$$

Under perfect recall (M. Zinkevich, 2007), we can write for any $h \in x$:

$$\rho^\pi(x) = \rho^{\pi^i}(h)\rho^{\pi^{-i}}(x)$$

And under perfect recall we will write for all $h \in x$, $\rho^{\pi^i}(x) = \rho^{\pi^i}(h)$ Furthermore, $V^i_\pi(h_{\text{init}}) = \sum_{h\in H}\rho^\pi(h)\sum_{a\in A}\pi(a|x(h))r^i_\pi(h,a)$

Value Function on the information states. The only information available to a player is the information state. We define the expected value of the game given such an information state x as follows:

$$V^i_\pi(x) = \frac{\sum_{h\in x}\rho^\pi(h)V^i_\pi(h)}{\sum_{h\in x}\rho^\pi(h)} \underbrace{=}_{\text{perfect recall}} \frac{\sum_{h\in x}\rho^{\pi^{-i}}(h)V^i_\pi(h)}{\sum_{h\in x}\rho^{\pi^{-i}}(h)}$$

And the expected Q-function given x and a is:

$$Q^i_\pi(x,a) = \frac{\sum_{h\in x}\rho^\pi(h)Q^i_\pi(h,a)}{\sum_{h\in x}\rho^\pi(h)} = \frac{\sum_{h\in x}\rho^{\pi^{-i}}(h)Q^i_\pi(h,a)}{\sum_{h\in x}\rho^{\pi^{-i}}(h)}$$

Now we can define a Nash equilibrium in the sequential imperfect information game setting. Formally:

Definition 3.1. A strategy π is a Nash equilibrium if for all $i \in \{1, \dots, N\}$ and for all π'^i: $V^i_{\pi'^i, \pi^{-i}}(h_{\text{init}}) \leq V^i_{\pi^i, \pi^{-i}}(h_{\text{init}})$

3.1. Monotone Games

In this paper, we are interested in: (i) zero-sum two-player games, i.e., $V^1_\pi = -V^2_\pi$ (many games implemented in OpenSpiel (Lanctot et al., 2019) fall in that category); (ii) in zero-sum N-player polymatrix games, i.e., when the value can be decomposed in a sum of pairwise interactions $V^i_\pi = \sum_{j \neq i} \tilde{V}^i_{\pi^i, \pi^j}$ with $\tilde{V}^i_{\pi^i, \pi^j} = -\tilde{V}^j_{\pi^j, \pi^i}$ generalizing Cai et al. (2016); Mertikopoulos et al. (2018); and finally (iii) in games where the profit of one player is decoupled from the interaction with the opponents, i.e., when the value can be decomposed in $V^i_\pi = \bar{V}^i_{\pi^i} + \bar{V}^i_{\pi^{-i}}$. All these settings can be captured by the following monotonicity condition:

Definition 3.2. Let us define $\Omega^i(\pi, \mu) = V^i_{\pi^i, \pi^{-i}}(h_{\text{init}}) - V^i_{\mu^i, \pi^{-i}}(h_{\text{init}}) - V^i_{\pi^i, \mu^{-i}}(h_{\text{init}}) + V^i_{\mu^i, \mu^{-i}}(h_{\text{init}})$. A game is monotone if for all policies π, μ, $\pi \neq \mu$:

$$\sum_{i \in \{1, \dots, N\}} \Omega^i(\pi, \mu) \leq 0$$

This condition is slightly difficult to interpret but as mentioned above, it captures a wide class of games (zero-sum two-player, polymatrix zero-sum games etc.). See proof in appendix I.

3.2. Follow the Regularized Leader

Follow the Regularized Leader in imperfect information games defines a sequence of policies $(\pi_s)_{s \geq 0}$ for all $i \in \{1, \dots, N\}$ and $x \in \mathcal{X}_i$ as follow:

$$y^i_t(x, a) = \int_0^t \rho^{\pi_s^{-i}}(x) Q^i_{\pi_s}(x, a) ds$$

$$\pi^i_t(.|x) = \operatorname*{arg\,max}_{p \in \Delta A} \Lambda^i(p, y^i_t(x, .))$$

We define the following quantity for any Nash equilibrium π^* of the game:

$$J(y) = \sum_{i=1}^N \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x) [\phi^*_i(y^i(x, .)) - \langle \pi^*(.|x), y^i(x, .) \rangle]$$

This quantity will be at the center of our analysis of Follow the Regularized Leader in sections 4 and 5. The following lemma shows how this quantity evolves if both players learn using Follow the Regularized Leader updates. We will use this quantity to create a Lyapunov function for policy-dependent reward and use it to bound the trajectories of FoReL to prove Poincaré recurrence; intuitively, that "most" trajectories do not converge to equilibria.

Lemma 3.1. *If y_t is defined as the follow the regularized leader dynamics we have:*

$$\frac{d}{dt} J(y) = \sum_{i=1}^N \underbrace{[V^i_{\pi^i_t, \pi^{*-i}} - V^i_{\pi^*}]}_{\leq 0} + \underbrace{\sum_{i=1}^N \Omega^i(\pi, \pi^*)}_{\leq 0 \text{ for a monotone game}}$$

$$+ \sum_{i=1}^N \sum_{h \in H \setminus \mathcal{Z}} \rho^{\pi_t^{-i}}(h) \rho^{\pi^{*i}}(h) \times$$

$$\mathbb{E}_{a \sim (\pi^{*i}, \pi_t^{-i})(..|x(h))} [r^i_{\pi^{*i}, \pi_t^{-i}}(h, a) - r^i_{\pi_t}(h, a)]$$

(proof in appendix A)

4. Recurrence of FoReL

This section generalizes the results of (Mertikopoulos et al., 2018) to Follow the Regularized Leader in monotone Imperfect Information Games when the reward is policy-independent (all the zero-sum two-player games implemented in OpenSpiel (Lanctot et al., 2019) have this property) and when the equilibrium has a full support. This requires two steps, first we will prove that an equivalent learning dynamic is **Divergence-free** (or preserves volume). Then we will use lemma 3.1 to show that all trajectories of this new dynamical system are **bounded**. This is enough to prove that the trajectories of FoReL are **Poincaré recurrent**. Intuitively this means that all trajectories will go back to a neighborhood of their starting point arbitrarily often. The Poincaré recurrence theorem (Piliouras & Shamma, 2014; Mertikopoulos et al., 2018; Poincaré, 1890) states:

Theorem 4.1. *If a flow preserves volume (is Divergence-free) and has only bounded orbits then for each open set there exist orbits that intersect the set infinitely often.*

Instead of studying the original dynamical system, we will fix an action a_x for all x and consider the dynamical system (as this system keeps w bounded):

$$\dot{w}^i_t(x, a) = \rho^{\pi_t^{-i}}(x)[Q^i_{\pi_t}(x, a) - Q^i_{\pi_t}(x, a_x)] \tag{1}$$

$$\pi^i_t(.|x) = \operatorname*{arg\,max}_{p \in \Delta A} \Lambda^i(p, w^i_t(x, .)) \tag{2}$$

Divergence-free. In order to get qualitative results on FoReL, we will prove that the FoReL dynamic is Divergence-free (a generalization of a result from Mertikopoulos et al. (2018)).

Lemma 4.1. *The system defined above (equation (1) and (2)) is autonomous (can be written as $\dot{w}_t = \xi(w_t)$), Divergence-free, and the dynamic of the policy π_t is equivalent to the one defined in section 3.2 when the reward is policy-independent. (Proof in appendix B)*

This property is critical as it implies that the dynamical system has no attractor (Weibull, 1997, p.252, prop 6.6).

Remark. This does not mean that the policy will not converge. If the w diverges, the policy might converge to a deterministic strategy. However, if the Nash is of full support, it will not be an attractor of the dynamical system.

We now know that Nash equilibria cannot be attractors of FoReL as the system is Divergence-free. In order to prove the Poincaré recurrence, we need to prove an additional property. We need the trajectory w_t to remain bounded if the equilibrium π^* is interior.

Lemma 4.2. *If the equilibrium is interior, then* $\sum_{i=1}^{N}[V^i_{\pi^i_t,\pi^{*-i}} - V^i_{\pi^*}] = 0.$

(Proof in appendix E)

Corollary 4.1. *In a monotone game with a policy-independent reward and an interior equilibrium, if y_t is defined as following the FoReL algorithm we have: $\frac{d}{dt}J(y) \leq 0$*

Corolary 4.1 implies that the trajectories of the dynamics equation (1) and (2) are bounded. This can be proven by directly using arguments from (Mertikopoulos et al., 2018, Lemma D.2.).

Poincaré recurrence. As we have seen in the two previous paragraphs, the flow of FoReL is Divergence-free and all trajectories are bounded in the case of monotone games with an interior Nash equilibrium. Thus all orbits are Poincaré recurrent.

5. Reward Transformation and Convergence in IIG

In section 4 we have seen that a policy-independent reward signal can lead to recurrent behavior. The idea we study here is to slightly modify the reward signal such that the Nash equilibrium of this new game is an attractor. We will show two reward transformations that guarantee convergence to a Nash equilibrium (with Lyapunov arguments). The first reward transformation applies generally to monotone games and the second one applies specifically to zero-sum games. But first we briefly recall the Lyapunov method.

Lyapunov method. The idea of the Lyapunov method to study the ordinary differential equation $\frac{d}{dt}y_t = \xi(y_t)$ is to look at the variations of a quantity $\mathcal{F}(y) \geq 0$ (and $\mathcal{F}(y^*) = 0$). The function $\mathcal{F}$ is said to be a strict Lyapunov function if:

$$\forall y \neq y^*,\ \frac{d}{dt}\mathcal{F}(y_t) < 0$$

In that case, the y_t will converge to a minimum of $\mathcal{F}$ if ξ is locally Lipschitz and if $\mathcal{F}$ is a continuously differentiable function. The function $\mathcal{F}$ is said to be a strong Lyapunov function if:

$$\frac{d}{dt}\mathcal{F}(y_t) \leq -\beta\mathcal{F}(y_t),\ \beta > 0$$

In this case, the y_t will converge to a minimum of $\mathcal{F}$ at an exponentially fast rate $\mathcal{F}(y_t) \leq \mathcal{F}(y_0)\exp(-\beta t)$.

Monotone games. In the general case of monotone games, the reward that for any μ preserves the monotonicity is: (see proof in section F)

$$r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log\frac{\pi(a|x(h))}{\mu(a|x(h))}$$

An immediate corollary of lemma 3.1 is:

Corollary 5.1. *In monotone games, the reward transformation $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log\frac{\pi(a|x(h))}{\mu(a|x(h))}$ considered above implies that J will be decreasing:*

$$\tfrac{d}{dt}J(y) \leq -\eta \textstyle\sum_{i=1}^{N}\sum_{h\in H_i} \rho^{\pi^{*i}}(h) KL(\pi^*(.|x(h)),\pi_t(.|x(h)))$$

Finally, if the regularizer ϕ_i is the entropy, we can show that the $\Xi(\pi^*,\pi_t) = \sum_{i=1}^{N}\sum_{h\in H_i} \rho^{\pi^{*i}}(h) KL(\pi^*(.|x(h)),\pi_t(.|x(h)))$ is a strong Lyapunov function:

Theorem 5.1. *If the regularizer ϕ_i is the entropy:*

$$\frac{d}{dt}\Xi(\pi^*,\pi_t) \leq -\eta\Xi(\pi^*,\pi_t)$$

it implies: $\Xi(\pi^*,\pi_t) \leq \Xi(\pi^*,\pi_0)\exp(-\eta t)$ (proof in appendix C)

This method thus introduces a trade-off between the speed of convergence of the algorithm and the transformation we make to the reward (which has an impact on the equilibrium of the transformed game).

Zero-sum two-player games. Whilst the above approach can be applied to all monotone games, the following reward can be applied specifically to zero-sum games. For any μ, this reward keeps the zero-sum property (see appendix F) and is more prone to sample based methods as the $\frac{1}{\rho^{\pi^{-i}}(h)}$ is not involved (with x being $x(h)$ and $\delta_{h,i} = \mathbf{1}_{i=\tau(h)}$),

$$r^i_\pi(h,a) = r^i(h,a) + \eta(1 - 2\delta_{h,i})\log\frac{\pi(a|x)}{\mu(a|x)}$$

And in that case:

Corollary 5.2. *We have:*

$$\frac{d}{dt}J(y) \leq$$
$$-\eta\sum_{i=1}^{N}\sum_{h\in H_i} \rho^{\pi^{*i}}(h)\rho^{\pi_t^{-i}}(h) KL(\pi^*(.|x(h)),\pi_t(.|x(h)))$$

And here, if the regularizer ϕ_i is the entropy, we can show that the $\Xi(\pi^*, \pi_t) = \sum_{i=1}^{N} \sum_{h \in H_i} \rho^{\pi^{*i}}(h) KL(\pi^*(.|x(h)), \pi_t(.|x(h)))$ is a strict Lyapunov function:

Theorem 5.2. *If the regularizer ϕ_i is the entropy:*

$$\frac{d}{dt}\Xi(\pi^*, \pi_t) \leq -\eta\zeta\Xi(\pi^*, \pi_t)$$

with $\zeta = \min_{x \in \mathcal{X}} \min_{\pi = \arg\max_p \Lambda(p,y) \text{ and } J(y) \leq J(y_0)} \sum_{h \in x} \rho^{\pi^{-i}}(h)$

This imply that: $\Xi(\pi^*, \pi_t) \leq \Xi(\pi^*, \pi_0)\exp(-\zeta\eta t)$

Proof. The proof follows by combining corollary 5.2 and the result in appendix C. □

In summary, we saw in this section that exponential convergence rates can be achieved in continuous time in imperfect information games using reward transformation.

Remark. Corrolary 5.1 and 5.2 are valid for all Nash of the transformed game. This means that for all η, the Nash eq. of the transformed game is unique. This uniqueness property is necessary to define the process of the next section.

6. Convergence to an Exact Equilibrium

The previous section introduced a reward transformation (by adding a policy dependent term $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$) to ensure exponential convergence in games. However this method does not ensure convergence to the equilibrium of the game defined on $r^i(h,a)$. In this section, we study the sequence of policies starting from π_0, being the uniform policy, and π_k the solution of the game with the reward transformation $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\pi_{k-1}(a|x(h))}$. Intuitively, this approach entails that the policy π_k will be searched close to the previous iterate π_{k-1} (we write $\pi_k = F(\pi_{k-1})$).

Lemma 6.1. *Then for any Nash equilibrium of the game π^*, we have the following identity for the sequence of policy π_k:*

$$\Xi(\pi^*, \pi_k) - \Xi(\pi^*, \pi_{k-1}) = -\Xi(\pi_k, \pi_{k-1}) + \frac{1}{\eta}\sum_{i=1}^{N}(m^i_k + \delta^i_k + \kappa^i_k)$$

Where:

$$\Xi(\mu, \pi) = \sum_{i=1}^{N} \sum_{h \in H_i} \rho^{\mu^i}(h) KL(\mu(.|x(h)), \pi(.|x(h)))$$

Where:

$$\kappa^i_k = \sum_{x \in \mathcal{X}^i} \rho^{\pi^{*i}}(x) \rho^{\pi_k^{-i}}(x) \times \sum_{a \in A} \left[\pi^{*i}(a|x(h)) - \pi_k(a|x(h))\right] \, {}^kQ^i_{\pi_k}(x,a) \leq 0$$

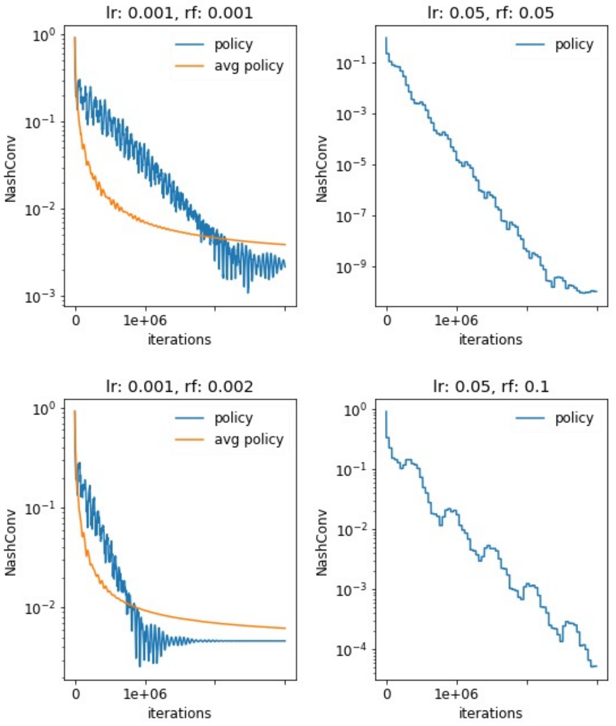

Figure 2. The left plots illustrate the monotone reward transform whilst the right plot illustrate the direct convergence method shown in section 6 where we change the reward every 40000 steps (rf is the value of the parameter η and lr is the time discretization).

Where: $\delta^i_k = V^i_{\pi^i_k, \pi^{*-i}}(h_{init}) - V^i_{\pi^*}(h_{init}) \leq 0$

And where:

$$m^i_k = V^i_{\pi_k}(h_{init}) - V^i_{\pi^{*i}, \pi_k^{-i}}(h_{init}) - V^i_{\pi^i_k, \pi^{*-i}}(h_{init}) + V^i_{\pi^*}(h_{init})$$

And where $\sum_{i=1}^{N} m^i_k \leq 0$ *if the game is monotone (proof in appendix D).*

Theorem 6.1. *In a monotone game with all Nash equilibrium being interior, the sequence of policy $\{\pi_k\}_{k \geq 0}$ (or $\{F^k(\pi_0)\}_{k \geq 0}$) converges to a Nash equilibrium of the game (proof in appendix H).*

Remark. We were only able to prove this result for interior Nash but we conjecture that it is still true for non interior Nash equilibrium.

7. Empirical evaluation

It has already been empirically noted that regularization helps convergence in games (Omidshafiei et al., 2019). Earlier work (Srinivasan et al., 2018) also provides experiments where the current policy converges in Leduc Poker, whilst the paper only proves convergence analysis of the average policy. Our work sheds a new light on those results as the convergence may have been the result of high regularization (the entropy cost added in (Srinivasan et al., 2018) appendix G was 0.1). The experiments will show how reward transform can be used to improve the state of the art of reinforcement Learning in Imperfect Information Games. To keep our implementation as close as possible to FoReL, we use the NeuRD policy update (Omidshafiei et al., 2019), a retrace update to estimate the Q-function. In order to keep our estimate of the return unbiased, we use that learned Q-function as a control variate as in (Schmid et al., 2019). The details of the algorithm are in appendix J. We present results on four games: Kuhn & Leduc Poker, Goofspiel and Liars Dice, which have respectively 12, 936, 162 and 24,576 information states. We evaluate all our policies using the NashConv metric (Lanctot et al., 2017) defined as $NashConv(\pi) = \sum_{i=1}^{N} \max_{\pi'^i} V^i_{\pi'^i,\pi^{-i}}(h_{\text{init}}) - V^i_{\pi}(h_{\text{init}})$.

In this section, we highlight two results with function approximation and illustrate the theory with tabular experiments on Kuhn Poker (figure 2). A more complete empirical evaluation and the precise description of the setting is available in appendix K.

7.1. Experiment with a Decaying Regularization

We found that decaying the regularization η exponentially from an initial value $\eta_{\max} = 1$ to a target value (we looked at values $\{1.0, 0.5, 0.2, 0.05, 0.01, 0.0\}$) is an effective empirical method. In figure 3 (top plot), we represent the NashConv as a function of the number of steps. We achieve our best performance for $\eta = 0.05$ with a NashConv of 0.10. This outperforms the results of NFSP (Heinrich & Silver, 2016), which has a best result of 0.12 in NashConv (0.06 of exploitability reported in the paper) and the state of the art algorithms implemented in Openspiel, which are no better than 0.2 in NashConv (this difference can be explaines by the fact that the implementation of Leduc Poker in the NFSP paper uses card isomorphism while in this paper we don't). However, for low choices of η the algorithm might diverge.

7.2. Iteration over the Regularization

As we have seen in section 6, the convergence to an exact equilibrium can be achieved by iteratively adapting the the reward. In the experiment (Fig. 3 bottom plot), we change the reward periodically every N-steps between steps $[kN, kN + \frac{N}{2}]$ we linearly interpolate between $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\pi_{kN}(a|x(h))}$ and $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\pi_{(k-1)N}(a|x(h))}$ and in interval $[kN + \frac{N}{2}, (k+1)N]$ we use the transformed reward $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\pi_{kN}(a|x(h))}$. As shown in Fig. 3 (bottom plot), this technique allows convergence for very high η. This is quite an advantage as the method will be more robust to the choice of that hyper-parameter.

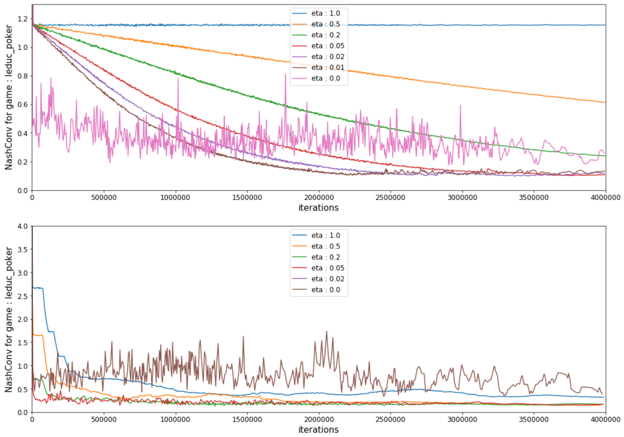

Figure 3. The precise setup used is described in appendix J. The top plot shows results on Leduc poker improving over the NFSP results using a decay of the regularization η. The bottom plot uses a fixed regularization and adapts the reward every $7.5e4$ steps as described in section 7.2.

	Leduc	Kuhn	Liars Dice	GoofSpie(4)
NFSP	0.16	0.02	0.25	0.14
Deep CFR	0.23	0.009	**0.19**	0.25
Q-learning	2.44	0.33	0.94	2.0
PSRO	0.17	**0.002**	0.28	0.23
NeuRD	**0.10**	0.02	0.25	**0.22**

Figure 4. Comparison of the best NashConv obtained for the best set of parameters between regularized NeuRD, NFSP, Deep CFR, Q-learning and PSRO. The sweep used is described in Appendix.L

8. Conclusion

We generalize the Poincaré recurrence result for FoReL from 2-player normal-form zero-sum games to sequential imperfect information games with a monotonicity condition. Although this is a generalization of a negative convergence result, we show that several reward transformations can guarantee convergence to a slightly modified equilibrium. We also show how to recover the original equilibrium of the game (when it is interior). Finally, based on these techniques we improve the state-of-the-art in model-free deep reinforcement learning in imperfect information games.

Since this work only focuses on FoReL, we aim to analyze the behavior of other dynamics in the sequential case from a dynamical systems perspective in future work. Fictitious play or softmax Q-learning have been theoretically considered in normal form games and their analysis with Lyapunov methods remains to be done in the IIG case. Furthermore, the role of regularization for convergence in games needs to be studied more systematically in other settings. Ideas like regularization could also be studied in for example Generative Adversarial Networks.

Acknowledgements

This research/project is supported in part by NRF2019-NRF-ANR095 ALIAS grant, grant PIE-SGP-AI-2018-01, NRF 2018 Fellowship NRF-NRFF2018-07, AME Programmatic Fund (Grant No. A20H6b0151) from the Agency for Science, Technology and Research (A*STAR) and the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG2-RP-2020-016). We would also like to thank Martin Schmid for the many reviews of this paper.

References

Abernethy, J., Lai, K. A., Levy, K. Y., and Wang, J.-K. Faster rates for convex-concave games. In *Conference on Learning Theory (COLT)*, 2018.

Bailey, J. P. and Piliouras, G. Multiplicative weights update in zero-sum games. In *ACM Conference on Economics and Computation*, 2018.

Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. The mechanics of n-player differentiable games. *International Conference on Machine Learning (ICML)*, 2018.

Bloembergen, D., Tuyls, K., Hennes, D., and Kaisers, M. Evolutionary dynamics of multi-agent learning: a survey. *Journal of Artificial Intelligence Research*, 53, 2015.

Boone, V. and Piliouras, G. From Darwin to Poincaré and von Neumann: Recurrence and cycles in evolutionary and algorithmic game theory. In *International Conference on Web and Internet Economics*, 2019.

Brown, G. W. Iterative solutions of games by fictitious play. In *Activity Analysis of Production and Allocation*, pp. 374–376, 1951.

Cai, Y., Candogan, O., Daskalakis, C., and Papadimitriou, C. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2): 648–655, 2016.

Cesa-Biachi, N. and Lugosi, G. *Predition, Learning, and Games*. Cambridge University Press, 2006.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *International Conference on Machine Learning (ICML)*, 2018.

Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2017.

Fudenberg, D. and Levine, D. *The Theory of Learning in Games*. MIT Press, 1998.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized Markov decision processes. In *International Conference on Machine Learning (ICML)*, 2019.

Gidel, G., Hemmat, R. A., Pezeshki, M., Huang, G., Lepriol, R., Lacoste-Julien, S., and Mitliagkas, I. Negative momentum for improved game dynamics. In *Artificial Intelligence and Statistics (AISTATS)*, 2019.

Gintis, H. *Game Theory Evolving*. Princeton University Press, 2nd edition, 2009.

Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., and Krause, A. An online learning approach to generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Heinrich, J. and Silver, D. Deep reinforcement learning from self-play in imperfect-information games. *arXiv*, 2016.

Heinrich, J., Lanctot, M., and Silver, D. Fictitious self-play in extensive-form games. In *International Conference on Machine Learning (ICML)*, 2015.

Hofbauer, J. and Sandholm, W. H. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6), 2002.

Hofbauer, J., Sorin, S., and Viossat, Y. Time average replicator and best-reply dynamics. *Mathematics of Operations Research*, 34(2):263–269, 2009.

Hu, J. and Wellman, M. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069, Nov 2003.

Kaisers, M. and Tuyls, K. Frequency adjusted multi-agent Q-learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010.

Kaisers, M. and Tuyls, K. FAQ-learning in matrix games: Demonstrating convergence near Nash equilibria, and bifurcation of attractors in the battle of sexes. In *AAAI Workshop on Interactive Decision Theory and Game Theory*, 2011.

Lagoudakis, M. G. and Parr, R. Value function approximation in zero-sum Markov games. In *Uncertainty in Artificial Intelligence (UAI)*, 2002.

Lanctot, M. *Monte Carlo Sampling and Regret Minimization for Equilibrium Computation and Decision-Making in Large Extensive Form Games*. PhD thesis, Department of Computing Science, University of Alberta, 2013.

Lanctot, M., Waugh, K., Zinkevich, M., and Bowling, M. Monte Carlo sampling for regret minimization in extensive games. In *Neural Information Processing Systems (NIPS)*, 2009.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning. In *Neural Information Processing Systems (NIPS)*, 2017.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., et al. Openspiel: A framework for reinforcement learning in games. *arXiv*, 2019.

Leslie, D. S. and Collins, E. J. Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2):495–514, 2005.

Letcher, A., Foerster, J., Balduzzi, D., Rocktäschel, T., and Whiteson, S. Stable opponent shaping in differentiable games. In *International Conference on Learning Representations (ICLR)*, 2019.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 1994.

Littman, M. L. Friend-or-foe Q-learning in general-sum games. In *International Conference on Machine Learning (ICML)*, 2001.

Lockhart, E., Lanctot, M., Pérolat, J., Lespiau, J.-B., Morrill, D., Timbers, F., and Tuyls, K. Computing approximate equilibria in sequential adversarial games by exploitability descent. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

M. Zinkevich, M. Bowling, N. B. A new algorithm for generating equilibria in massive zero-sum games. In *AAAI Conference on Artificial Intelligence*, 2007.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. Independent reinforcement learners in cooperative markov games: A survey regarding coordination problems. *Knowl. Eng. Rev.*, 27(1):1–31, February 2012.

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

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. Cycles in adversarial regularized learning. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2018.

Mescheder, L., Nowozin, S., and Geiger, A. The numerics of GANs. In *Neural Information Processing Systems (NIPS)*, 2017.

Oliehoek, F. A., Savani, R., Gallego-Posada, J., Van der Pol, E., De Jong, E. D., and Groß, R. GANGs: Generative adversarial network games. *arXiv*, 2017.

Omidshafiei, S., Hennes, D., Morrill, D., Munos, R., Perolat, J., Lanctot, M., Gruslys, A., Lespiau, J.-B., and Tuyls, K. Neural replicator dynamics. *arXiv*, 2019.

Ortega, P. A. and Legg, S. Modeling friends and foes. *arXiv*, 2018.

Pérolat, J., Scherrer, B., Piot, B., and Pietquin, O. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning (ICML)*, 2015.

Pérolat, J., Piot, B., Geist, M., Scherrer, B., and Pietquin, O. Softened approximate policy iteration for Markov games. In *International Conference on Machine Learning (ICML)*, 2016.

Pérolat, J., Piot, B., Scherrer, B., and Pietquin, O. On the use of non-stationary strategies for solving two-player zero-sum Markov games. In *Artificial Intelligence and Statistics (AISTATS)*, 2016.

Pérolat, J., Strub, F., Piot, B., and Pietquin, O. Learning Nash equilibrium for general-sum Markov games from batch data. *Artificial Intelligence and Statistics (AISTATS)*, 2017.

Piliouras, G. and Shamma, J. S. Optimization despite chaos: Convex relaxations to complex limit sets via Poincaré recurrence. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2014.

Poincaré, H. Sur le problème des trois corps et les équations de la dynamique. *Acta mathematica*, 13(1):A3–A270, 1890.

Schmid, M., Burch, N., Lanctot, M., Moravcik, M., Kadlec, R., and Bowling, M. Variance reduction in Monte Carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. In *AAAI Conference on Artificial Intelligence*, 2019.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Singh, S. P., Kearns, M. J., and Mansour, Y. Nash convergence of gradient dynamics in general-sum games. In *Uncertainty in Artificial Intelligence (UAI)*, 2000.

Srinivasan, S., Lanctot, M., Zambaldi, V., Pérolat, J., Tuyls, K., Munos, R., and Bowling, M. Actor-critic policy optimization in partially observable multiagent environments. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. Fast convergence of regularized learning in games. In *Neural Information Processing Systems (NIPS)*, 2015.

Taylor and Jonker. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40:145–156, 1978.

Tuyls, K. and Nowé, A. Evolutionary game theory and multi-agent reinforcement learning. *The Knowledge Engineering Review*, 20(1):63–90, 2005.

Tuyls, K., Verbeeck, K., and Lenaerts, T. A selection-mutation model for Q-learning in multi-agent systems. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2003.

Waugh, K. and Bagnell, J. A. A unified view of large-scale zero-sum equilibrium computation. In *AAAI Conference on Artificial Intelligence Workshops*, 2015.

Weibull, J. Evolutionary game theory. *MIT press*, 1997.

Zeeman, E. Population dynamics from game theory. *Lecture Notes in Mathematics, Global theory of dynamical systems*, 819, 1980.

Zeeman, E. Dynamics of the evolution of animal conflicts. *Theoretical Biology*, 89:249–270, 1981.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. In *Neural Information Processing Systems (NIPS)*, 2008.

A. Proof of Lemma 3.1

$$
\begin{aligned}
\frac{d}{dt}J(y) &= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}_i} \rho^{\pi^{*i}}(x)\rho^{\pi_t^{-i}}(x)\langle \pi_t(.|x) - \pi^*(.|x), Q^i_{\pi_t}(x,.)\rangle \\
&= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}_i} \rho^{\pi^{*i}}(x) \sum_{h\in x} \rho^{\pi_t^{-i}}(h)\langle \pi_t(.|x(h)) - \pi^*(.|x(h)), Q^i_{\pi_t}(h,.)\rangle \\
&= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}_i} \sum_{h\in x} \rho^{\pi_t^{-i}}(h)\rho^{\pi^{*i}}(h)\langle \pi_t(.|x(h)) - \pi^*(.|x(h)), Q^i_{\pi_t}(h,.)\rangle \\
&= \sum_{i=1}^{N} \sum_{h\in H_i} \rho^{\pi_t^{-i}}(h)\rho^{\pi^{*i}}(h)\langle \pi_t(.|x(h)) - \pi^*(.|x(h)), Q^i_{\pi_t}(h,.)\rangle
\end{aligned}
$$

Let's write ${}^i\bar{\pi}_t = (\pi^{*i}, \pi_t^{-i})$ and let's notice that for all $h \in H^{-i}$, we have that ${}^i\bar{\pi}_t(.|x(h)) = \pi_t(.|x(h))$ and thus for all $h \in H^{-i}$, $\langle \pi_t(.|x(h)) - {}^i\bar{\pi}_t(.|x(h)), Q^i_{\pi_t}(h,.)\rangle = 0$

$$
\begin{aligned}
\frac{d}{dt}J(y) &= \sum_{i=1}^{N} \sum_{h\in H} \rho^{\pi_t^{-i}}(h)\rho^{\pi^{*i}}(h)\langle \pi_t(.|x(h)) - {}^i\bar{\pi}_t(.|x(h)), Q^i_{\pi_t}(h,.)\rangle \\
&= \sum_{i=1}^{N} \left[\sum_{h\in H\setminus\mathcal{Z}} \rho^{{}^i\bar{\pi}_t}(h) \left[V^i_{\pi_t}(h) - \sum_{a\in A} {}^i\bar{\pi}_t(a|x(h))(r^i_{\pi_t}(h,a) + V^i_{\pi_t}(ha)) \right] + \sum_{h\in\mathcal{Z}} \rho^{{}^i\bar{\pi}_t}(h) V^i_{\pi_t}(h) \right] \\
&= \left[\sum_{i=1}^{N} \sum_{h\in H} \rho^{{}^i\bar{\pi}_t}(h) V^i_{\pi_t}(h) \right] - \left[\sum_{i=1}^{N} \sum_{h\in H\setminus\{h_{\text{init}}\}} \rho^{{}^i\bar{\pi}_t}(h) V^i_{\pi_t}(h) \right] - \left[\sum_{i=1}^{N} \sum_{h\in H\setminus\mathcal{Z}} \rho^{{}^i\bar{\pi}_t}(h) \sum_{a\in A} {}^i\bar{\pi}_t(a|x(h)) r^i_{\pi_t}(h,a) \right] \\
&= \sum_{i=1}^{N} V^i_{\pi_t}(h_{\text{init}}) - \sum_{i=1}^{N} \sum_{h\in H\setminus\mathcal{Z}} \rho^{{}^i\bar{\pi}_t}(h) \sum_{a\in A} {}^i\bar{\pi}_t(a|x(h)) r^i_{{}^i\bar{\pi}_t}(h,a) \\
&\qquad + \sum_{i=1}^{N} \sum_{h\in H\setminus\mathcal{Z}} \rho^{{}^i\bar{\pi}_t}(h) \sum_{a\in A} {}^i\bar{\pi}_t(a|x(h)) [r^i_{{}^i\bar{\pi}_t}(h,a) - r^i_{\pi_t}(h,a)] \\
&= \left[\sum_{i=1}^{N} V^i_{\pi_t}(h_{\text{init}}) - V^i_{(\pi^{*i}, \pi_t^{-i})}(h_{\text{init}}) \right] + \sum_{i=1}^{N} \sum_{h\in H\setminus\mathcal{Z}} \rho^{\pi_t^{-i}}(h)\rho^{\pi^{*i}}(h) \mathbb{E}_{a\sim(\pi^{*i},\pi_t^{-i})(..|x(h))}[r^i_{\pi^{*i},\pi_t^{-i}}(h,a) - r^i_{\pi_t}(h,a)] \\
&= \left[\sum_{i=1}^{N} V^i_{(\pi_t^i, \pi^{*-i})}(h_{\text{init}}) - V^i_{\pi^*}(h_{\text{init}}) \right] + \left[\sum_{i=1}^{N} V^i_{\pi_t}(h_{\text{init}}) - V^i_{(\pi^{*i}, \pi_t^{-i})}(h_{\text{init}}) - V^i_{(\pi_t^i, \pi^{*-i})}(h_{\text{init}}) + V^i_{\pi^*}(h_{\text{init}}) \right] \\
&+ \sum_{i=1}^{N} \sum_{h\in H\setminus\mathcal{Z}} \rho^{\pi_t^{-i}}(h)\rho^{\pi^{*i}}(h) \mathbb{E}_{a\sim(\pi^{*i},\pi_t^{-i})(..|x(h))}[r^i_{\pi^{*i},\pi_t^{-i}}(h,a) - r^i_{\pi_t}(h,a)]
\end{aligned}
$$

Which concludes the proof.

B. The system is equivalent to FoReL dynamics and is Divergence-free (lemma 4.1)

The dynamical system

$$\begin{aligned} \dot{w}_t^i(x,a) &= \rho^{\pi_t^{-i}}(x)[Q^i_{\pi_t}(x,a) - Q^i_{\pi_t}(x,a_x)] \\ \pi_t^i(.|x) &= \arg\max_{p\in\Delta A} \Lambda^i(p, w_t^i(x,.)) \end{aligned}$$

And

$$\begin{aligned} \dot{y}_t^i(x,a) &= \rho^{\tilde{\pi}_t^{-i}}(x) Q^i_{\tilde{\pi}_t}(x,a) \\ \tilde{\pi}_t^i(.|x) &= \arg\max_{p\in\Delta A} \Lambda^i(p, y_t^i(x,.)) \end{aligned}$$

generate the same sequence of policies

Proof. For all $x \in \mathcal{X}_i$ the variable:

$$y_t^i(x,a) = \int_{s=0}^{t} \rho^{\tilde{\pi}_s^{-i}}(x) Q^i_{\tilde{\pi}_s}(x,a) ds$$

and we define $\tilde{y}_t^i(x,a)$ as:

$$\begin{aligned} &\tilde{y}_t^i(x,a) \\ &= y_t^i(x,a) - y_t^i(x,a_x) \\ &= \int_{s=0}^{t} \rho^{\tilde{\pi}_s^{-i}}(x) Q^i_{\tilde{\pi}_s}(x,a) ds - \int_{s=0}^{t} \rho^{\tilde{\pi}_s^{-i}}(x) Q^i_{\tilde{\pi}_s}(x,a_x) ds \\ &= \int_{s=0}^{t} \rho^{\tilde{\pi}_s^{-i}}(x) \left[Q^i_{\tilde{\pi}_s}(x,a) - Q^i_{\tilde{\pi}_s}(x,a_x)\right] ds \end{aligned}$$

$$\tilde{\pi}_t^i(.|x) = \arg\max_{p\in\Delta A} \Lambda^i(p, y_t^i(x,.)) = \arg\max_{p\in\Delta A} = \Lambda^i(p, y_t^i(x,.) - y_t^i(x,a_x)) = \Lambda^i(p, \tilde{y}_t^i(x,a))$$

Thus $\tilde{y}_t^i(x,a)$ and $y_t^i(x,a)$ generate the same sequence of policy.

And since $\tilde{y}_t^i(x,a)$ and $w_t^i(x,a)$ follow the same differential equation and have the same initial conditions, $w_t^i(x,a)$ and $y_t^i(x,a)$ generate the same sequence of policies. □

The dynamical system:

$$\begin{aligned} \dot{w}_t^i(x,a) &= \rho^{\pi_t^{-i}}(x)[Q^i_{\pi_t}(x,a) - Q^i_{\pi_t}(x,a_x)] \\ \pi_t^i(.|x) &= \arg\max_{p\in\Delta A} \Lambda^i(p, w_t^i(x,.)) \end{aligned}$$

is an autonomous dynamical system as π_t^i is a function of $w_t^i(x,a)$. Let us write it $w_t = \xi(w_t)$ we have $\xi(w_t)(i,x,a) = \rho^{\pi_t^{-i}}(x)[Q^i_{\pi_t}(x,a) - Q^i_{\pi_t}(x,a_x)]$.

Finally, $\forall i \in \{1,\dots,N\}, \forall x \in \mathcal{X}_i, \forall a \in A$, $\xi(w)(i,x,a) = \rho^{\pi^{-i}}(x)[Q^i_\pi(x,a) - Q^i_\pi(x,a_x)]$ (where $\pi^i(.|x) = \arg\max_{p\in\Delta A} \Lambda^i(p, w^i(x,.)))$ is independent of $w^i(x,a)$ as $\rho^{\pi^{-i}}(x) Q^i_\pi(x,a) = \sum\limits_{h\in x} [r^i(h,a) + V^i_\pi(ha)]$ does not depend on $\pi^i(.|x)$.

Thus we have $\frac{\partial \xi(w)(i,x,a)}{\partial w^i(x,a)} = 0$. This proves that the $div_w \xi(w) = \sum\limits_{i=1}^{N} \sum\limits_{x\in\mathcal{X}_i} \sum\limits_{a\in A} \frac{\partial \xi(w)(i,x,a)}{\partial w^i(x,a)} = 0$ and that the dynamics is incompressible.

C. Proof Strong Lyapunov Function

$$
\begin{aligned}
J(y) + \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)\phi_i(\pi^*(.|x)) &= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)[\phi_i^*(y^i(x,.)) - \langle \pi^*(.|x), y^i(x,.)\rangle + \phi_i(\pi^*(.|x))] \\
= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)[\Lambda^i(\pi^i(.|x), y^i(x,.)) - \langle \pi^*(.|x), y^i(x,.)\rangle + \phi_i(\pi^*(.|x))] \\
= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)[\langle \pi^i(.|x), y^i(x,.)\rangle - \phi_i(\pi^i(.|x)) - \langle \pi^*(.|x), y^i(x,.)\rangle + \phi_i(\pi^*(.|x))] \\
= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)[\phi_i(\pi^*(.|x)) - \phi_i(\pi^i(.|x)) + \langle \pi^i(.|x) - \pi^*(.|x), y^i(x,.)\rangle]
\end{aligned}
$$

for all y^i in $\{y^i \mid \sum_{a^i \in A^i} y^i(a^i) = 0\}$, the tangent space of ΔA^i, we have that $\nabla h_i(\nabla h_i^*(y^i)) = y^i$ if $\nabla h_i^*(y^i)$ is in the interior of ΔA^i (see (Hofbauer & Sandholm, 2002) for the statement of this property). Thus, for all y^i there exists a δ such that $\nabla h_i(\nabla h_i^*(y^i)) = y^i + \delta \mathbf{1}$

In the end:

$$
\begin{aligned}
J(y) + \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)\phi_i(\pi^*(.|x)) &= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)[\phi_i(\pi^*(.|x)) - \phi_i(\pi^i(.|x)) + \langle \pi^i(.|x) - \pi^*(.|x), y^i(x,.)\rangle] \\
= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)[\phi_i(\pi^*(.|x)) - \phi_i(\pi^i(.|x)) + \langle \pi^i(.|x) - \pi^*(.|x), \nabla\phi_i(\nabla\phi_i^*(y^i(x,.)))\rangle] \\
= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)[\phi_i(\pi^*(.|x)) - \phi_i(\pi^i(.|x)) + \langle \pi^i(.|x) - \pi^*(.|x), \nabla\phi_i(\pi^i(.|x))\rangle] \\
= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x) D_{\phi_i}(\pi^*(.|x), \pi^i(.|x))
\end{aligned}
$$

Where D_{ϕ_i} is the Bregman divergence associated with ϕ_i. If ϕ_i is the entropy, we the following equality $J(y) + \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x)\phi_i(\pi^*(.|x)) = \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*i}}(x) KL(\pi^*(.|x), \pi^i(.|x)) = \Xi(\pi^*, \pi)$.

That is why:

$$
\frac{d}{dt} J(y) = \frac{d}{dt} \Xi(\pi^*, \pi)
$$

D. Proof of Lemma 6.1

Let us write ${}^k r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\pi_{k-1}(a|x(h))}$ and ${}^i\bar{\pi}_k = (\pi^{*i}, \pi_k^{-i})$

$$\underbrace{\sum_{h\in H} \rho^{\pi_k}(h) \sum_{a\in A} \pi_k(a|x(h))^k r^i_{\pi_k}(h,a)}_{(2)} = \underbrace{\sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \sum_{a\in A} {}^i\bar{\pi}_k(a|x(h))\ {}^k r^i_{\pi_k}(h,a)}_{(3)}$$

$$+ \underbrace{\sum_{h\in H} \rho^{\pi_k}(h) \sum_{a\in A} \pi_k(a|x(h))^k r^i_{\pi_k}(h,a) - \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \sum_{a\in A} {}^i\bar{\pi}_k(a|x(h))\ {}^k r^i_{\pi_k}(h,a)}_{(1)}$$

Let us write the value function for the reward ${}^k r^i_\pi(h,a)$ and policy π_k will be written ${}^k V^i_{\pi_k}(h) = \sum_a \pi_k(a|x(h)) \left[{}^k r^i_{\pi_k}(h,a) + {}^k V^i_{\pi_k}(ha)\right]$

$$\begin{aligned}
(1) &= {}^k V^i_{\pi_k}(h_{\text{init}}) - \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \sum_{a\in A} {}^i\bar{\pi}_k(a|x(h))\ {}^k r^i_{\pi_k}(h,a) \\
&= \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h)\ {}^k V^i_{\pi_k}(h) - \sum_{h\in H\setminus\{h_{\text{init}}\}} \rho^{{}^i\bar{\pi}_k}(h)\ {}^k V^i_{\pi_k}(h) - \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \sum_{a\in A} {}^i\bar{\pi}_k(a|x(h))\ {}^k r^i_{\pi_k}(h,a) \\
&= \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \left[{}^k V^i_{\pi_k}(h) - \sum_{a\in A} {}^i\bar{\pi}_k(a|x(h))\ {}^k V^i_{\pi_k}(ha) \right] - \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \sum_{a\in A} {}^i\bar{\pi}_k(a|x(h))\ {}^k r^i_{\pi_k}(h,a) \\
&= \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \left[{}^k V^i_{\pi_k}(h) - \sum_{a\in A} {}^i\bar{\pi}_k(a|x(h)) \left[{}^k r^i_{\pi_k}(h,a) +\ {}^k V^i_{\pi_k}(ha)\right] \right] \\
&= \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \sum_{a\in A} \left[\pi_k(a|x(h)) - {}^i\bar{\pi}_k(a|x(h))\right]\ {}^k \\
&= \sum_{x\in\mathcal{X}^i} \sum_{h\in x} \rho^{\pi^{*i}}(h) \rho^{\pi_k^{-i}}(h) \sum_{a\in A} \left[\pi_k(a|x(h)) - \pi^{*i}(a|x(h))\right]\ {}^k Q^i_{\pi_k}(h,a) \\
&= \sum_{x\in\mathcal{X}^i} \rho^{\pi^{*i}}(x) \underbrace{\sum_{h\in x} \rho^{\pi_k^{-i}}(h)}_{\rho^{\pi_k^{-i}}(x)} \sum_{a\in A} \left[\pi_k(a|x(h)) - \pi^{*i}(a|x(h))\right]\ {}^k Q^i_{\pi_k}(h,a) \text{ from perfect recall} \\
&= \sum_{x\in\mathcal{X}^i} \rho^{\pi^{*i}}(x) \left[\sum_{h\in x} \rho^{\pi_k^{-i}}(h)\right] \sum_{a\in A} \left[\pi_k(a|x(h)) - \pi^{*i}(a|x(h))\right] \underbrace{\frac{\sum\limits_{h\in x} \rho^{\pi_k^{-i}}(h)\ {}^k Q^i_{\pi_k}(h,a)}{\sum\limits_{h\in x} \rho^{\pi_k^{-i}}(h)}}_{{}^k Q^i_{\pi_k}(x,a)} \\
&= \sum_{x\in\mathcal{X}^i} \rho^{\pi^{*i}}(x) \rho^{\pi_k^{-i}}(x) \sum_{a\in A} \left[\pi_k(a|x(h)) - \pi^{*i}(a|x(h))\right]\ {}^k Q^i_{\pi_k}(x,a) \\
&= -\underbrace{\kappa^i_k}_{\leq 0} \geq 0 \text{ as } \pi_k \text{ is a Nash for the game defined on reward } {}^k r^i_\pi(h,a)
\end{aligned}$$

Then:

$$\begin{aligned}
(2) &= \sum_{h\in H} \rho^{\pi_k}(h) \sum_{a\in A} \pi_k(a|x(h)) r^i(h,a) - \eta \sum_{h\in H^i} \rho^{\pi^i_k}(h) \sum_{a\in A} \pi^i_k(a|x(h)) \log \frac{\pi^i_k(a|x(h))}{\pi^i_{k-1}(a|x(h))} \\
&= V^i_{\pi_k}(h_{\text{init}}) - \eta \sum_{h\in H^i} \rho^{\pi^i_k}(h) KL\left(\pi^i_k(.|x(h)), \pi^i_{k-1}(.|x(h))\right)
\end{aligned}$$

And Finally:

$$\begin{aligned}(3) &= \sum_{h\in H} \rho^{{}^i\bar{\pi}_k}(h) \sum_{a\in A} {}^i\bar{\pi}_k(a|x(h)) r^i(h,a) - \eta \sum_{h\in H^i} \rho^{\pi^{*i}}(h) \sum_{a\in A} \pi^{*i}(a|x(h)) \log \frac{\pi_k(a|x(h))}{\pi_{k-1}(a|x(h))} \\ &= V^i_{\pi^{*i},\pi^{-i}_k}(h_{\text{init}}) - \eta \sum_{h\in H^i} \rho^{\pi^{*i}}(h) \left[KL\left(\pi^{*i}(.|x(h)), \pi_{k-1}(.|x(h))\right) - KL\left(\pi^{*i}(.|x(h)), \pi_k(.|x(h))\right)\right]\end{aligned}$$

Now combining (2) = (3) + (1) we have:

$$\begin{aligned}&\eta \sum_{h\in H^i} \rho^{\pi^{*i}}(h) KL\left(\pi^{*i}(.|x(h)), \pi_k(.|x(h))\right) - \eta \sum_{h\in H^i} \rho^{\pi^{*i}}(h) KL\left(\pi^{*i}(.|x(h)), \pi_{k-1}(.|x(h))\right) \\ &= V^i_{\pi_k}(h_{\text{init}}) - V^i_{\pi^{*i},\pi^{-i}_k}(h_{\text{init}}) + \kappa^i_k - \eta \sum_{h\in H^i} \rho^{\pi^i_k}(h) KL\left(\pi^i_k(.|x(h)), \pi^i_{k-1}(.|x(h))\right) \\ &= \underbrace{V^i_{\pi_k}(h_{\text{init}}) - V^i_{\pi^{*i},\pi^{-i}_k}(h_{\text{init}}) - V^i_{\pi^i_k,\pi^{*-i}}(h_{\text{init}}) + V^i_{\pi^*}(h_{\text{init}})}_{=m^i_k} + \underbrace{V^i_{\pi^i_k,\pi^{*-i}}(h_{\text{init}}) - V^i_{\pi^*}(h_{\text{init}})}_{=\delta^i_k} + \kappa^i_k \\ &\quad - \eta \sum_{h\in H^i} \rho^{\pi^i_k}(h) KL\left(\pi^i_k(.|x(h)), \pi^i_{k-1}(.|x(h))\right)\end{aligned}$$

And finally we have the desired property:

$$\begin{aligned}&\sum_{h\in H^i} \rho^{\pi^{*i}}(h) KL\left(\pi^{*i}(.|x(h)), \pi_k(.|x(h))\right) - \sum_{h\in H^i} \rho^{\pi^{*i}}(h) KL\left(\pi^{*i}(.|x(h)), \pi_{k-1}(.|x(h))\right) \\ =& \qquad \frac{1}{\eta} m^i_k + \frac{1}{\eta}\delta^i_k + \frac{1}{\eta}\kappa^i_k - \sum_{h\in H^i} \rho^{\pi^i_k}(h) KL\left(\pi^i_k(.|x(h)), \pi^i_{k-1}(.|x(h))\right)\end{aligned}$$

And we get the result by summing over the players:

$$\begin{aligned}&\sum_{i=1}^N \sum_{h\in H^i} \rho^{\pi^{*i}}(h) KL\left(\pi^{*i}(.|x(h)), \pi_k(.|x(h))\right) - \sum_{i=1}^N \sum_{h\in H^i} \rho^{\pi^{*i}}(h) KL\left(\pi^{*i}(.|x(h)), \pi_{k-1}(.|x(h))\right) \\ =& \qquad \frac{1}{\eta} \sum_{i=1}^N m^i_k + \frac{1}{\eta} \sum_{i=1}^N \delta^i_k + \frac{1}{\eta} \sum_{i=1}^N \kappa^i_k - \sum_{i=1}^N \sum_{h\in H^i} \rho^{\pi^i_k}(h) KL\left(\pi^i_k(.|x(h)), \pi^i_{k-1}(.|x(h))\right)\end{aligned}$$

E. Proof of Lemma 4.2

If the equilibrium is interior, then $\sum_{i=1}^{N} [V^i_{\pi^i_t,\pi^{*-i}} - V^i_{\pi^*}] = 0$.

Proof. First let us show that $\forall i, \pi^i$:

$$
\begin{aligned}
&V^i_{\pi^i,\pi^{*-i}} - V^i_{\pi^*} && (3)\\
&= \sum_{x \in \mathcal{X}_i} \rho^{\pi^i}(x) \rho^{\pi^{*-i}}(x) \sum_{a \in A} \left({\pi^*}^i(a|x) - \pi^i(a|x) \right) Q^i_{\pi^*}(x,a) && (4)
\end{aligned}
$$

Since π^* is a Nash equilibrium we always have $V^i_{\pi^i,\pi^{*-i}} - V^i_{\pi^*} \leq 0$. Let us suppose that there exists an information state x such that $Q^i_{\pi^*}(x,a)$ does not have the same values for all actions and that the equilibrium is of full support. Then a greedy policy on that state x (and π^* on the other states) should improve the value for player i. This would contradict π^* being a Nash equilibrium. This proves that all Q-values $Q^i_{\pi^*}(x,a)$ are equals for every states x. Then $\sum_{a \in A} \left({\pi^*}^i(a|x) - \pi^i(a|x) \right) Q^i_{\pi^*}(x,a) = 0$ for all states.

This conclud the proof that for all t, $\sum_{i=1}^{N} [V^i_{\pi^i_t,\pi^{*-i}} - V^i_{\pi^*}] = 0$. □

F. Reward Transformation in Monotone Games

The reward transformation that can be considered are the following:

$$r^i_\pi(h,a) = r^i(h,a) - \mathbf{1}_{i=\tau(h)} \eta \frac{\log \pi(a|x(h))}{\rho^{\pi^{-i}}(h)}$$

or,

$$r^i_\pi(h,a) = r^i(h,a) - \mathbf{1}_{i=\tau(h)} \eta \frac{\log \pi(a|x(h))}{\rho^{\pi^{-i}}(x(h))}$$

or for any μ,

$$r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$$

or finally for any μ,

$$r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(x(h))} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$$

And in that case we have:

$$\tfrac{d}{dt} J(y) = \textstyle\sum_{i=1}^N [V^i_{\pi^i_t, \pi^{*-i}} - V^i_{\pi^*}] + \sum_{i=1}^N \Omega^i(\pi, \pi^*) - \eta \sum_{i=1}^N \sum_{h \in H_i} \rho^{\pi^{*i}}(h) KL(\pi^*(.|x(h)), \pi_t(.|x(h)))$$

Proof. For the game defined on reward $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$

$$\begin{aligned}
V^i_\pi(h_{\text{init}}) &= \sum_{h \in H} \rho^\pi(h) \sum_{a \in A} \pi(a|x(h)) r^i_\pi(h,a) \\
&= \sum_{h \in H} \rho^\pi(h) \sum_{a \in A} \pi(a|x(h)) \left[r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))} \right] \\
&= \sum_{h \in H} \rho^\pi(h) \sum_{a \in A} \pi(a|x(h)) r^i(h,a) - \underbrace{\eta \sum_{h \in H^i} \rho^{\pi^i}(h) \sum_{a \in A} \pi^i(a|x(h)) \log \frac{\pi^i(a|x(h))}{\mu^i(a|x(h))}}_{\text{Only depends on } \pi^i.}
\end{aligned}$$

Thus if $\sum_{i=1}^N \Omega^i(\pi, \pi^*) = 0$ for the game defined with reward $r^i(h,a)$, then $\sum_{i=1}^N \Omega^i(\pi, \pi^*) = 0$ for the game defined on reward $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$ the monotonicity is also $\sum_{i=1}^N \Omega^i(\pi, \pi^*) = 0$.

$$\begin{aligned}
&\sum_{i=1}^N \sum_{h \in H \setminus \mathcal{Z}} \rho^{\pi_t^{-i}}(h) \rho^{\pi^{*i}}(h) \mathbb{E}_{a \sim (\pi^{*i}, \pi_t^{-i})(..|x(h))} [r^i_{\pi^{*i}, \pi_t^{-i}}(h,a) - r^i_{\pi_t}(h,a)] \\
&= -\eta \sum_{i=1}^N \sum_{h \in H^i \setminus \mathcal{Z}} \rho^{\pi^{*i}}(h) \sum_{a \in A} \pi^{*i}(a|x(h)) \log \frac{\pi^{*i}(a|x(h))}{\pi^i(a|x(h))}
\end{aligned}$$

The other cases are left in appendix. □

G. Reward Transformation in Zero-Sum Games

$$r^i_\pi(h,a) = r^i(h,a) - \mathbf{1}_{i=\tau(h)}\eta\log\pi(a|x(h)) + \mathbf{1}_{i\neq\tau(h)}\eta\log\pi(a|x(h)) \tag{5}$$

or for any μ,

$$r^i_\pi(h,a) = r^i(h,a) - \mathbf{1}_{i=\tau(h)}\eta\log\frac{\pi(a|x(h))}{\mu(a|x(h))} + \mathbf{1}_{i\neq\tau(h)}\eta\log\frac{\pi(a|x(h))}{\mu(a|x(h))} \tag{6}$$

And in that case:

$$\frac{d}{dt}J(y) = \sum_{i=1}^N[V^i_{\pi^i_t,\pi^{*-i}} - V^i_{\pi^*}] - \eta\sum_{i=1}^N\sum_{h\in H_i}\rho^{\pi^{*i}}(h)\rho^{\pi_t{}^{-i}}(h)KL(\pi^*(.|x(h)),\pi_t(.|x(h)))$$

Proof. The game is still zero-sum so the monotonicity is still zero.

$$\begin{aligned}&\sum_{i=1}^N\sum_{h\in H\backslash\mathcal{Z}}\rho^{\pi_t^{-i}}(h)\rho^{\pi^{*i}}(h)\mathbb{E}_{a\sim(\pi^{*i},\pi_t^{-i})(..|x(h))}[r^i_{\pi^{*i},\pi_t^{-i}}(h,a) - r^i_{\pi_t}(h,a)]\\ &= -\eta\sum_{i=1}^N\sum_{h\in H^i\backslash\mathcal{Z}}\rho^{\pi^{*i}}(h)\rho^{\pi_t^{-i}}(h)\sum_{a\in A}\pi^{*i}(a|x(h))\log\frac{\pi^{*i}(a|x(h))}{\pi^i(a|x(h))}\end{aligned}$$

As in that case $r^i_{\pi^{*i},\pi_t^{-i}}(h,a) - r^i_{\pi_t}(h,a) = 0$ for all $h\in H^{-i}$.

□

This regularization term increases the monotonicity of the game and thus helps the convergence of the method. This introduces however a bias in the Nash. Thus this method makes a bias-convergence trade off.

H. Convergence to a Nash

The proof of the convergence to an exact Nash uses similar arguments used to prove convergence for strict Lyapunov functions in the discrete vase. From lemma 6.1 we know that for a policy sequence starting from π_0 being the uniform policy and π_k is the solution of the game with the reward transformation $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\pi_{k-1}(a|x(h))}$. In this section, we will call this map F (and $F(\mu) = \pi^*_\mu$ is the equilibrium of the game defined on $r^i_\pi(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$). We will show that $\pi_k = F^k(\pi_0)$ will converge to a Nash equilibrium of the game π^*.

The proof proceeds in 3 steps:

- First we prove that F is continuous,
- Second we prove that $\min_{\pi^* \in \Pi^*} \Xi(\pi^*, F(\mu)) - \min_{\pi^* \in \Pi^*} \Xi(\pi^*, \mu) < 0$,
- The second step is enough to prove that $\min_{\pi^* \in \Pi^*} \Xi(\pi^*, \pi_k)$ converges to a value c. The last step proves by contradiction that c can't be anything but 0.

H.1. Continuity

The first step is to show that the map $F(.)$ which associate μ to the Nash equilibrium over the game defined over $r^i_{\mu,\pi}(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$ is continuous.

Then for all μ, μ', we have $r^i_{\mu,\pi}(h,a) - r^i_{\mu',\pi}(h,a) = -\frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\mu'(a|x(h))}{\mu(a|x(h))}$

Let us write now w^*_μ and $w^*_{\mu'}$ the fixed point of the dynamic defined in lemma 4.1 and Ξ_μ and $\Xi_{\mu'}$ their corresponding Lyapunov function and π^*_μ and $\pi^*_{\mu'}$.

Let us consider that $\tilde{w}$ follow the following ODE (where $Q^i_{\mu,\pi_t}(x,a)$ is the Q-function for reward $r^i_{\mu,\pi}(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$)

$$\dot{\tilde{w}}^i_t(x,a) = \rho^{\pi_t^{-i}}(x)[Q^i_{\mu,\pi_t}(x,a) - Q^i_{\mu,\pi_t}(x,a_x)]$$
$$\pi^i_t(.|x) = \operatorname*{arg\,max}_{p \in \Delta A} \Lambda^i(p, \tilde{w}^i_t(x,.))$$

Let us suppose furthermore that we start from the equilibrium $\tilde{w}(0) = w^*_\mu$

Let us examine the variation of $\Xi_{\mu'}(\tilde{w}(t))$ and write $\pi_t(.|x) = \Gamma(\tilde{w}(t)(.,x))$:

$$\begin{aligned}
\frac{d}{dt}\Xi(\pi^*_{\mu'}, \pi_t) &= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*}_{\mu'}{}^{i}}(x) \rho^{\pi_t^{-i}}(x) \langle \pi_t(.|x) - \pi^*_{\mu'}(.|x), Q^i_{\mu,\pi_t}(x,.) \rangle \\
&= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*}_{\mu'}{}^{i}}(x) \rho^{\pi_t^{-i}}(x) \langle \pi_t(.|x) - \pi^*_{\mu'}(.|x), Q^i_{\mu',\pi_t}(x,.) \rangle \\
&\quad + \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_i} \rho^{\pi^{*}_{\mu'}{}^{i}}(x) \rho^{\pi_t^{-i}}(x) \langle \pi_t(.|x) - \pi^*_{\mu'}(.|x), Q^i_{\mu,\pi_t}(x,.) - Q^i_{\mu',\pi_t}(x,.) \rangle
\end{aligned}$$

Let have $\mu, \mu' \in D_0$ (where D_0 is an open set). Furthermore let us suppose that for all $\mu \in D_0$ there exists $\epsilon > 0$ such that for all $x \in \mathcal{X}$ and $a \in A$ $\mu(a|x) > \epsilon$.

The function $\log$ is locally Lipschitz of constant K in D_0.

As $\pi_t = \pi^*_\mu$ we can bound $Q^i_{\mu,\pi_t}(x,.) - Q^i_{\mu',\pi_t}(x,.) \leq \eta T_{\max} \left[\sup_{\mu'' \in D_0} \max_{h \in H_i} \frac{1}{\rho^{\pi^{*}_{\mu''}{}^{-i}}(h)} \right] K \|\mu - \mu'\|$

Finally We can have that:

$$\frac{d}{dt}\Xi(\pi^*_{\mu'}, \pi_t) \leq -\eta\Xi(\pi^*_{\mu'}, \pi_t) + \eta T_{\max}\left[\sup_{\mu'' \in D_0} \max_{h \in H_i} \frac{1}{\rho^{\pi^{*-i}_{\mu''}}(h)}\right] K\|\mu - \mu'\|$$

This imply that $\Xi(\pi^*_{\mu'}, \pi^*_{\mu}) \leq T_{\max}\left[\sup_{\mu'' \in D_0} \max_{h \in H_i} \frac{1}{\rho^{\pi^{*-i}_{\mu''}}(h)}\right] K\|\mu - \mu'\|$.

This finally imply that the map $\mu \to \pi^*_{\mu}$ is continuous.

H.2. The function $\min_{\pi* \in \Pi^*} \Xi(\pi^*, \mu)$ is a strict Lyapunov function

We have seen that the following equality holds (in lemma 6.1):

$\Xi(\pi^*, \pi_k) - \Xi(\pi^*, \pi_{k-1}) = -\Xi(\pi_k, \pi_{k-1}) + \frac{1}{\eta}\sum_{i=1}^{N} m^i_k + \frac{1}{\eta}\sum_{i=1}^{N} \delta^i_k + \frac{1}{\eta}\sum_{i=1}^{N} \kappa^i_k$

Where:

$$\Xi(\mu, \pi) = \sum_{i=1}^{N} \sum_{h \in H_i} \rho^{\mu^i}(h) KL(\mu(.|x(h)), \pi(.|x(h)))$$

Where:

$$\kappa^i_k = \sum_{x \in \mathcal{X}^i} \rho^{\pi^{*i}}(x)\rho^{\pi^{-i}_k}(x) \sum_{a \in A} \left[\pi^{*i}(a|x(h)) - \pi_k(a|x(h))\right] \, {}^kQ^i_{\pi_k}(x, a) \leq 0$$

Where:

$$\delta^i_k = V^i_{\pi^i_k, \pi^{*-i}}(h_{\text{init}}) - V^i_{\pi^*}(h_{\text{init}}) \leq 0$$

And where:

$$m^i_k = V^i_{\pi_k}(h_{\text{init}}) - V^i_{\pi^{*i}, \pi^{-i}_k}(h_{\text{init}}) - V^i_{\pi^i_k, \pi^{*-i}}(h_{\text{init}}) + V^i_{\pi^*}(h_{\text{init}})$$

And where $\sum_{i=1}^{N} m^i_k \leq 0$ if the game is monotone.

First let us write Π^* the set of Nash equilibrium if the game defined on reward $r^i(h, a)$.

The Goal of this section is to prove that that $\min_{\pi^* \in \Pi^*} \Xi(\pi^*, F(\mu)) - \min_{\pi^* \in \Pi^*} \Xi(\pi^*, \mu) < 0$ **for all** $\mu \notin \Pi^*$

The first step of our proof is to show if there exists a k such that $\Xi(F(\mu), \mu) = 0$, then $F(\mu), \mu \in \Pi^*$.

To do so, we first need to prove a serie of technical lemma.

Lemma H.1. *For all π, π^* and for all $i \in \{1, \ldots, N\}$ we have:*

$$V^i_{\pi^*}(h_{\textit{init}}) - V^i_{\pi^i, \pi^{*-i}}(h_{\textit{init}}) = \sum_{x \in \mathcal{X}_i} \rho^{\pi^i}(x)\rho^{\pi^{*-i}}(x) \sum_{a} (\pi^i(a|x) - \pi^{*i}(a|x))Q^i_{\pi^*}(x, a)$$

Proof. Let's write $\bar{\pi} = (\pi^i, \pi^{*-i})$

$$\begin{aligned}
& V^i_{\pi^*}(h_{\text{init}}) - V^i_{\pi^i,\pi^{*-i}}(h_{\text{init}}) \\
& = \underbrace{\sum_{h \in H} \rho^{\pi^i}(h)\rho^{\pi^{*-i}}(h)[V^i_{\pi^*}(h) - \sum_{a \in A} \bar{\pi}(a|x(h))V^i_{\pi^*}(ha)]}_{=V^i_{\pi^*}(h_{\text{init}})} - \underbrace{\sum_{h \in H} \rho^{\pi^i}(h)\rho^{\pi^{*-i}}(h) \sum_{a \in A} \bar{\pi}(a|x(h))r^i(h,a)]}_{=V^i_{\pi^i,\pi^{*-i}}(h_{\text{init}})} \\
& = \sum_{h \in H} \rho^{\pi^i}(h)\rho^{\pi^{*-i}}(h)[V^i_{\pi^*}(h) - \sum_{a \in A} \bar{\pi}(a|x(h))[r^i(h,a) + V^i_{\pi^*}(ha)]] \\
& = \sum_{h \in H} \rho^{\pi^i}(h)\rho^{\pi^{*-i}}(h) \sum_{a \in A} (\pi^*(a|x(h)) - \bar{\pi}(a|x(h)))[Q^i_{\pi^*}(h,a)] \\
& = \sum_{h \in H_i} \rho^{\pi^i}(h)\rho^{\pi^{*-i}}(h) \sum_{a \in A} (\pi^{*i}(a|x(h)) - \pi^i(a|x(h)))Q^i_{\pi^*}(h,a) \text{ as } \bar{\pi} = \pi* \text{ on all the opponent nodes.} \\
& = \sum_{x \in \mathcal{X}_i} \rho^{\pi^i}(x)\rho^{\pi^{*-i}}(x) \sum_{a \in A} (\pi^{*i}(a|x) - \pi^i(a|x))[Q^i_{\pi^*}(x,a)]
\end{aligned}$$

□

Lemma H.2. *Let π^* be a policy. If for all $i \in \{1, \ldots, N\}, x \in \mathcal{X}_i$ and $\hat{\pi}$ such that $\sum_a (\pi^*(a|x) - \hat{\pi}(a|x))\, Q^i_{\pi^*}(x,a) \geq 0$ then π^* is a Nash equilibrium.*

Proof. Let suppose that for all $i \in \{1, \ldots, N\}, x \in \mathcal{X}_i$ and $\hat{\pi}$ we have $\sum_a (\pi^*(a|x) - \hat{\pi}(a|x))\, Q^i_{\pi^*}(x,a) \geq 0$

Then by lemma H.1 we have:

$$\forall i,\ V^i_{\pi^*}(h_{\text{init}}) - V^i_{\pi^i,\pi^{*-i}}(h_{\text{init}}) = \sum_{x \in \mathcal{X}_i} \rho^{\pi^i}(x)\rho^{\pi^{*-i}}(x) \sum_a (\pi^i(a|x) - \pi^{*i}(a|x))Q^i_{\pi^*}(x,a) \geq 0$$

Thus π^* is a Nash equilibrium. □

Corollary H.1. *If π^* is not a Nash equilibrium, then there exists $i \in \{1, \ldots, N\}, x \in \mathcal{X}_i$ and $\hat{\pi}$ such that $\sum_a (\pi^*(a|x) - \hat{\pi}(a|x))\, Q^i_{\pi^*}(x,a) < 0$*

Proof. This is a direct consequence of lemma H.2 □

Theorem H.1. *If $\pi^*_\mu = F(\mu) = \mu$, then μ is a Nash equilibrium of the game defined on $r^i(h,a)$.*

Proof. First we will write $V^i_{\mu,\pi}(h)$ ($Q^i_{\mu,\pi}(h,a)$) to be the value function (Q-function) with respect to the reward $r^i_{\mu,\pi}(h,a) = r^i(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{-i}}(h)} \log \frac{\pi(a|x(h))}{\mu(a|x(h))}$.

The reader will notice that since $\pi^*_\mu = \mu$ then $Q^i_{\mu,\pi^*_\mu}(h,a) = Q^i_{\pi^*_\mu}(h,a)$.

We will prove the result by contradiction. Let suppose that π^*_μ is not a Nash equilibrium for the game with reward $r^i(h,a)$. Then there exists i, $\hat{\pi}$ and $\tilde{x} \in \mathcal{X}_i$ such that $\sum_a \left(\pi^*_\mu(a|\tilde{x}) - \hat{\pi}(a|\tilde{x})\right) Q^i_{\pi^*_\mu}(\tilde{x},a) < 0$

For the rest of this proof, we will write $\hat{\pi}_\alpha$ the policy defined as π^*_μ on all $x \in \mathcal{X} \backslash \{\tilde{x}\}$ and $(1-\alpha)\pi^*_\mu + \alpha\hat{\pi}$ on state $\tilde{x}$.

As π^*_μ is a Nash equilibrium for the reward $r^i_{\mu,\pi}(h,a)$, then $V^i_{\mu,\pi^*_\mu}(h_{\text{init}}) - V^i_{\mu,\hat{\pi}_\alpha}(h_{\text{init}}) \geq 0$

$$\begin{aligned}
&V^i_{\mu,\pi^*_\mu}(h_{\text{init}}) - V^i_{\mu,\hat{\pi}_\alpha}(h_{\text{init}}) \\
&= \sum_{h\in H} \rho^{\hat{\pi}_\alpha}(h)[V^i_{\mu,\pi^*_\mu}(h) - \sum_{a\in A} \hat{\pi}_\alpha(a|x(h))V^i_{\mu,\pi^*_\mu}(ha)] - \sum_{h\in H} \rho^{\hat{\pi}_\alpha}(h) \sum_{a\in A} \hat{\pi}_\alpha(a|x(h)) r^i_{\mu,\hat{\pi}_\alpha}(h,a)] \\
&= \sum_{h\in \tilde{x}} \rho^{\hat{\pi}_\alpha}(h)[V^i_{\mu,\pi^*_\mu}(h) - \sum_{a\in A} \hat{\pi}_\alpha(a|x(h))[r^i_{\mu,\hat{\pi}_\alpha}(h,a) + V^i_{\mu,\pi^*_\mu}(ha)]] \text{ as } \hat{\pi}_\alpha \text{ and } \pi^*_\mu \text{ are only different on } \tilde{x} \\
&= \sum_{h\in \tilde{x}} \rho^{\pi^*_\mu}(h)[V^i_{\mu,\pi^*_\mu}(h) - \sum_{a\in A} \hat{\pi}_\alpha(a|x(h))[Q^i_{\mu,\pi^*_\mu}(h,a) - \frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{*-i}_\mu}(h)} \log \frac{\hat{\pi}_\alpha(a|x(h))}{\mu(a|x(h))}]] \\
&= \sum_{h\in \tilde{x}} \rho^{\pi^*_\mu}(h)[\frac{\eta \mathbf{1}_{i=\tau(h)}}{\rho^{\pi^{*-i}_\mu}(h)} KL(\hat{\pi}_\alpha(.|x(h)), \mu(.|x(h))) + \sum_{a\in A} [\pi^*_\mu(a|x(h)) - \hat{\pi}_\alpha(a|x(h))] Q^i_{\mu,\pi^*_\mu}(h,a)] \\
&= \eta \left(\sum_{h\in \tilde{x}} \rho^{\pi^{*i}_\mu}(h) \right) KL(\hat{\pi}_\alpha(.|\tilde{x}), \mu(.|\tilde{x})) + \rho^{\pi^*_\mu}(\tilde{x}) \sum_{a\in A} [\pi^*_\mu(a|\tilde{x}) - \hat{\pi}_\alpha(a|\tilde{x})] Q^i_{\mu,\pi^*_\mu}(\tilde{x},a)] \\
&\leq \eta \left(\sum_{h\in \tilde{x}} \rho^{\pi^{*i}_\mu}(h) \right) \frac{1}{2} \|\hat{\pi}_\alpha(.|\tilde{x}) - \mu(.|\tilde{x})\|_1^2 + \rho^{\pi^*_\mu}(\tilde{x}) \sum_{a\in A} [\pi^*_\mu(a|\tilde{x}) - \hat{\pi}_\alpha(a|\tilde{x})] Q^i_{\mu,\pi^*_\mu}(\tilde{x},a)] \text{ by the Pinsker inequality.} \\
&\leq \eta\alpha^2 \left(\sum_{h\in \tilde{x}} \rho^{\pi^{*i}_\mu}(h) \right) \frac{1}{2} \|\hat{\pi}(.|\tilde{x}) - \mu(.|\tilde{x})\|_1^2 + \alpha\rho^{\pi^*_\mu}(\tilde{x}) \underbrace{\sum_{a\in A} [\pi^*_\mu(a|\tilde{x}) - \hat{\pi}(a|\tilde{x})] Q^i_{\mu,\pi^*_\mu}(\tilde{x},a)]}_{<0 \text{ as } Q^i_{\mu,\pi^*_\mu}(\tilde{x},a) = Q^i_{\pi^*_\mu}(\tilde{x},a)}
\end{aligned}$$

So there exists $c > 0$ and $d < 0$ such that $V^i_{\mu,\pi^*_\mu}(h_{\text{init}}) - V^i_{\mu,\hat{\pi}_\alpha}(h_{\text{init}}) \leq c\alpha^2 + d\alpha = c\alpha(\alpha + \frac{d}{c})$. And finally there exists $\alpha > 0$ such that $V^i_{\mu,\pi^*_\mu}(h_{\text{init}}) - V^i_{\mu,\hat{\pi}_\alpha}(h_{\text{init}}) < 0$ which contradicts the fact that π^*_μ is a Nash for the game on reward $r^i_{\mu,\pi}(h,a)$.

From theorem H.1 we can conclude that for all $\mu \notin \Pi^*$, we have $\Xi(F(\mu), \mu) > 0$ this directly imply that for all $\mu \notin \Pi^*$ we have $\min_{\pi^*\in\Pi^*} \Xi(\pi^*, F(\mu)) - \min_{\pi^*\in\Pi^*} \Xi(\pi^*, \mu) < 0$.

□

H.3. Convergence to the Nash

We want to show that the sequence of policies $\pi_k = F^k(\pi_0)$ converges to a Nash equilibrium of the game. And we suppose that all policies π^* are interior.

Under these conditions, we have the following properties:

- $F(.)$ is a continuous map on the interior of the simplex (see section H.1),
- for all $\mu \notin \Pi^*$ we have $\min_{\pi^*\in\Pi^*} \Xi(\pi^*, F(\mu)) - \min_{\pi^*\in\Pi^*} \Xi(\pi^*, \mu) < 0$.
- $\mu \rightarrow \min_{\pi^*\in\Pi^*} \Xi(\pi^*, F(\mu))$ is a positive function infinite on the border of the simplex continuous in μ.
- $\mu \rightarrow \Delta V(\mu) = \min_{\pi^*\in\Pi^*} \Xi(\pi^*, F(\mu)) - \min_{\pi^*\in\Pi^*} \Xi(\pi^*, \mu)$ is continuous in μ.

Let us write $\bar{\Omega}_c = \{\mu| \ \min_{\pi^*\in\Pi^*} \Xi(\pi^*, \mu) \leq c\}$ and $\Omega_c = \{\mu| \ \min_{\pi^*\in\Pi^*} \Xi(\pi^*, \mu) \leq c\}$. For all finite c the set Ω_c is closed and bounded set (and $\bar{\Omega}_c$ is an open bounded set). Then Ω_c is a compact set.

Let us consider that $C = \min_{\pi^*\in\Pi^*} \Xi(\pi^*, \pi_0)$. Since $\Delta V(\mu) < 0$ then the sequence of $\min_{\pi^*\in\Pi^*} \Xi(\pi^*, \pi_k)$ converges to c.

By contradiction let us suppose that $c > 0$. This means that all the π_k are all in the closed set $\Omega_{C,c} = \bar{\Omega}_C \backslash \Omega_c$. The set $\Omega_{C,c}$ is bounded and thus is a compact. The image of $\Omega_{C,c}$ through $\Delta V(.)$ (which is a continuous map) is a compact K and $k_{\max} = \sup_{x\in K} < 0$. This means that $\forall k, \Delta V(\pi_k) \leq k_{\max}$

This means that

$$\begin{aligned}\min_{\pi^*\in\Pi^*}\Xi(\pi^*,\pi_k) &\leq \min_{\pi^*\in\Pi^*}\Xi(\pi^*,\pi_0)+\sum_{i=0}^{k-1}\Delta V(\pi_i)\\ &\leq C+k\times k_{\max}\end{aligned}$$

This contradicts the fact that $c>0$ as there exists k such that $C+k\times k_{\max}<c$.

In the end $k\rightarrow\min_{\pi^*\in\Pi^*}\Xi(\pi^*,\pi_k)$ converges to 0 and π_k converges to Π^*

I. Monotone Games

In this section, we prove that zero-sum, zero-sum N-player polymatrix games and games where the profit of one player is decoupled from the interaction with the opponent are monotone.

(i) zero-sum two-player games, i.e., $V^1_\pi = -V^2_\pi$, the monotonicity condition becomes:

$\Omega^i(\pi, \mu) = V^i_{\pi^i,\pi^{-i}}(h_{\text{init}}) - V^i_{\mu^i,\pi^{-i}}(h_{\text{init}}) - V^i_{\pi^i,\mu^{-i}}(h_{\text{init}}) + V^i_{\mu^i,\mu^{-i}}(h_{\text{init}})$

It is easy to notice that $\Omega^1(\pi, \mu) = -\Omega^2(\pi, \mu)$ as $\forall \pi, \mu \; V^1_{\mu^1,\pi^2}(h_{\text{init}}) = -V^2_{\mu^1,\pi^2}(h_{\text{init}})$

(ii) in zero-sum N-player polymatrix games, i.e., when the value can be decomposed in a sum of pairwise interactions $V^i_\pi = \sum\limits_{j \neq i} \tilde{V}^i_{\pi^i,\pi^j}$ with $\tilde{V}^i_{\pi^i,\pi^j} = -\tilde{V}^j_{\pi^j,\pi^i}$

In that case:

$$\begin{aligned}
\sum_{i \in \{1,\ldots,N\}} \Omega^i(\pi, \mu) = &\sum_{i \in \{1,\ldots,N\}} \sum_{j \neq i} [\tilde{V}^i_{\pi^i,\pi^j}(h_{\text{init}}) - V^i_{\pi^i,\mu^j}(h_{\text{init}}) - V^i_{\mu^i,\pi^j}(h_{\text{init}}) + V^i_{\mu^i,\mu^j}(h_{\text{init}})] \\
&\sum_{i \in \{1,\ldots,N\}} \sum_{j < i} [\tilde{V}^i_{\pi^i,\pi^j}(h_{\text{init}}) - V^i_{\pi^i,\mu^j}(h_{\text{init}}) - V^i_{\mu^i,\pi^j}(h_{\text{init}}) + V^i_{\mu^i,\mu^j}(h_{\text{init}})] \\
&\qquad + [\tilde{V}^j_{\pi^j,\pi^i}(h_{\text{init}}) - V^j_{\pi^j,\mu^i}(h_{\text{init}}) - V^j_{\mu^j,\pi^i}(h_{\text{init}}) + V^j_{\mu^j,\mu^i}(h_{\text{init}})] \\
= \sum_{i \in \{1,\ldots,N\}} \sum_{j < i} 0 = 0 &
\end{aligned}$$

(iii) in games where the profit of one player is decoupled from the interaction with the opponents, i.e., when the value can be decomposed in $V^i_\pi = \bar{V}^i_{\pi^i} + \bar{V}^i_{\pi^{-i}}$.

In this case $\Omega^i(\pi, \mu) = V^i_{\pi^i,\pi^{-i}}(h_{\text{init}}) - V^i_{\mu^i,\pi^{-i}}(h_{\text{init}}) - V^i_{\pi^i,\mu^{-i}}(h_{\text{init}}) + V^i_{\mu^i,\mu^{-i}}(h_{\text{init}}) = \bar{V}^i_{\pi^i} + \bar{V}^i_{\pi^{-i}} - [\bar{V}^i_{\mu^i} + \bar{V}^i_{\pi^{-i}}] - [\bar{V}^i_{\pi^i} + \bar{V}^i_{\mu^{-i}}] + \bar{V}^i_{\mu^i} + \bar{V}^i_{\mu^{-i}} = 0$

J. Empirical set up

batch size	the batch size used to update the policy and the Q-function
batch size actor	to fasten the experiment we use an actor critic setup with 512 actors all of them produces batch of trajectories.
eta	the parameter used to transform the reward
lambda Retrace	the parameter of retrace
epsilon greedy	the epsilon greedy policy parameter
Gradient clipping value	all gradient on the policy and the value are clipped
max number of steps	the maximum number of steps
reward re-centered every	the recentering periode
policy learning rate start	the policy learning rate starts its exponential decay at this value
policy learning rate end	the exponential decay ends at this value
threshold NeuRD	the NeuRD threshold
Neural net structure for π and Q	MLP with 2 hidden layer of 128 unit

Figure 5. This table summarize the meaning of all hyperparameters of the algorithm

J.1. Estimation of the Critique

The update on a Q-function is done such as to minimize the l_2-norm between $\hat{Q}^i_{\boldsymbol{w}}(x_l, a_l)$ a retrace target (Espeholt et al., 2018) constructed using the sequence or policies and rewards (written $Q^i_{\text{retrace target}}$).

$$w_i \leftarrow w_i + \alpha \sum_{l=0}^{K} \tau(x_l) \times \left[Q^i_{\text{retrace target}}(x_l, a_l) - \hat{Q}^i_{\boldsymbol{w}}(x_l, a_l)\right] \partial_{w_i} \hat{Q}^i_{\boldsymbol{w}}(x_l, a_l). \tag{7}$$

J.2. Low Variance Unbiased Estimate of the Expected Payoff

This version of NeuRD uses an unbiased estimate and low-variance of the return (Schmid et al., 2019). We will account that the policy $\pi^i_{\theta_i}(a^i|x_l)$ we want to evaluate can be different from the one we are sampling $\nu^i_{\theta_i}(a^i|x_l)$ and the unbiased return is computed as follow in the case of the reward transform for zero-sum games as follows:

$$\begin{aligned}
&\bar{Q}^i_{\boldsymbol{w}}(x_l, a) \\
&= \left\{ \begin{array}{ll} -\eta \log(\pi(a|x_l)) + \hat{Q}^i_{\boldsymbol{w}}(x_l, a) & \text{if } a \neq a_l \\ \\ -\eta \log(\pi(a|x_l)) + \hat{Q}^i_{\boldsymbol{w}}(x_l, a) & \\ +\frac{1}{\nu(a^i|x_l)} \left[r^i(x_l, a_l) + E_{b \sim \pi(.|x_{l+1})}[\bar{Q}^i_{\boldsymbol{w}}(x_{l+1}, b)] - \hat{Q}^i_{\boldsymbol{w}}(x_l, a)\right] & \text{if } a = a_l \end{array} \right\}
\end{aligned}$$

if $\tau(x_l) = i$

$$\bar{Q}^i_{\boldsymbol{w}}(x_l, a) = \left\{ \begin{array}{ll} \eta \log(\pi(a|x_l)) & \text{if } a \neq a_l \\ \eta \log(\pi(a|x_l)) + \frac{1}{\nu(a^i|x_l)} \left[r^i(x_l, a_l) + E_{b \sim \pi(.|x_{l+1})}[\bar{Q}^i_{\boldsymbol{w}}(x_{l+1}, b)]\right] & \text{if } a = a_l \end{array} \right\}$$

if $\tau(x_l) \neq i$

By convention, we will have that $\forall a, \bar{Q}^i_{\boldsymbol{w}}(x_{K+1}, a) = 0$

J.3. NeuRD update

In the second step we correct for the reach probability:

$$\tilde{Q}^i_{\boldsymbol{w}}(x_l, a) = \left(\prod_{k=0}^{l-1} \frac{\mathbf{1}_{\tau x_l \neq i} \pi(a_k|x_l) + \mathbf{1}_{\tau x_l = i}}{\nu(a_k|x_l)} \right) \bar{Q}^i_{\boldsymbol{w}}(x_l, a)$$

Where π and ν are the policies at the player's turn.

The policy update follows the following equation:

$$\theta_i \leftarrow \theta_i + \alpha \sum_{l=0}^{K} \mathbf{1}_{\tau x_l = i} \sum_{a^i} \partial_{\theta_i} \xi^i_{\theta_i}(a^i|x_l) \left[\tilde{Q}^i_{\boldsymbol{w}}(x_l, a^i) \right]. \tag{8}$$

Where $\xi^i_{\theta_i}$ is the logit of policy and $\text{softmax}(\xi^i_{\theta_i}) = \pi^i_{\theta_i}$. The NeuRD update rule require an additional clipping parameter to avoid numerical instabilities. We leave the reader to (Omidshafiei et al., 2019).

Last, we obtained the exploration policy ν by doing an epsilon greedy policy π.

K. Experiments

K.1. Tabular Experiments

In this section we present experiments on Kuhn poker and Leduc poker that illustrate the convergence property for the dynamics on the transformed reward.

K.1.1. Tabular Experiments With a Fixed Regularization (reward transformation for zero-sum games)

The two following figures illustrate the method described in section 5. The following experiment Shows the FoReL dynamics on Kuhn poker :

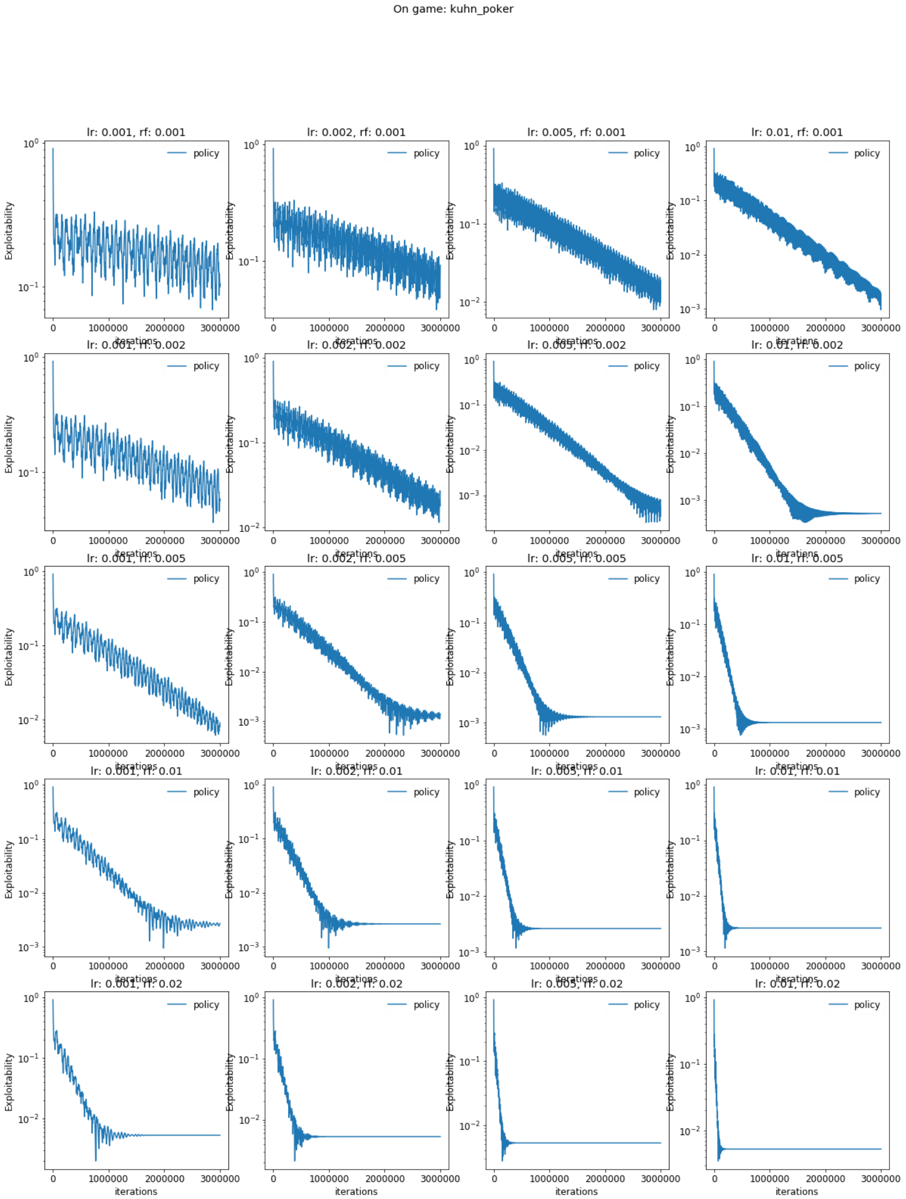

Figure 6. Kuhn Poker.

And the following experiment Shows the FoReL dynamics on Leduc poker :

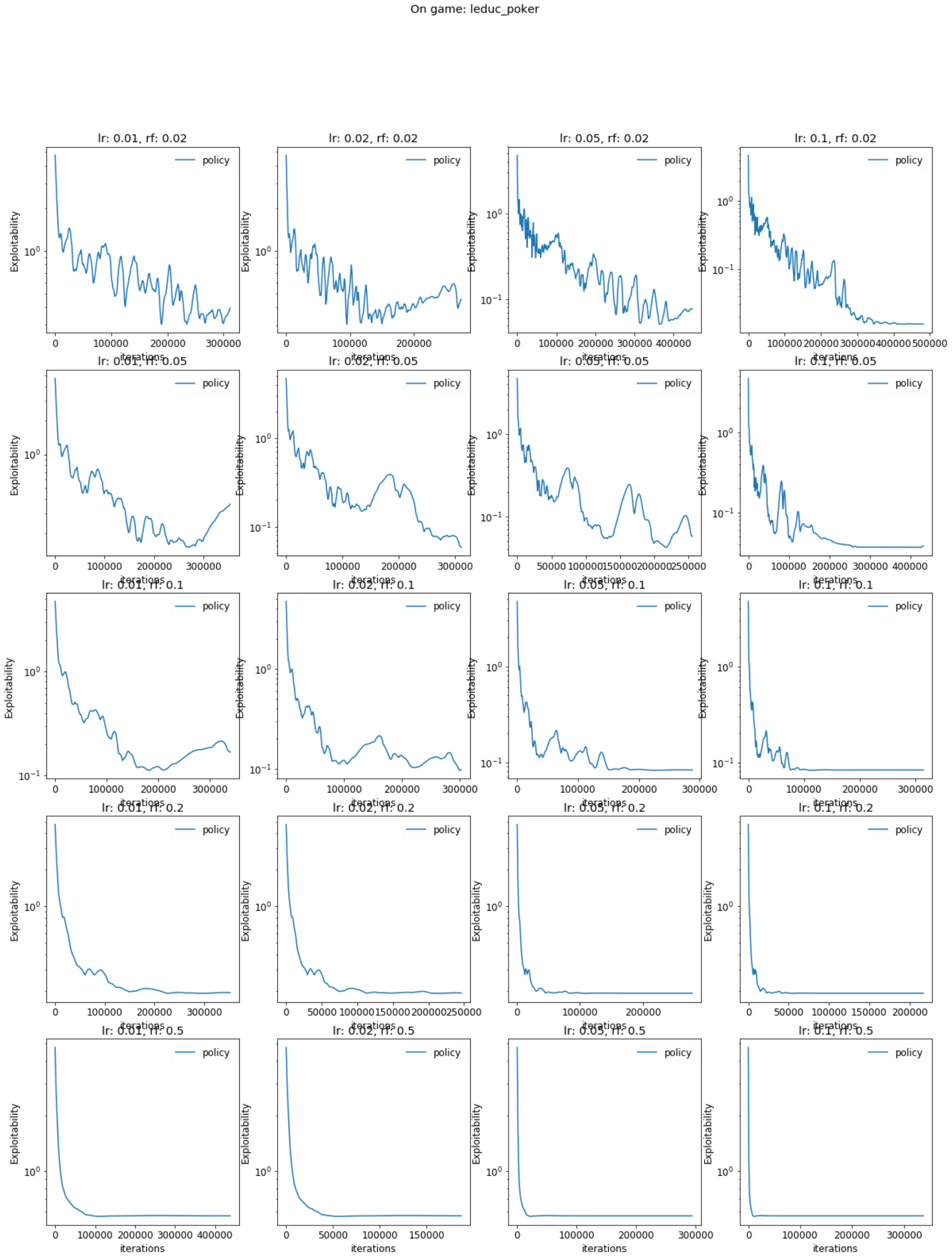

Figure 7. Leduc Poker.

K.1.2. Tabular Experiments With a Fixed Regularization (reward transformation for monotone games)

The two following figures illustrate the method described in section 5. The following experiment Shows the FoReL dynamics on Kuhn poker :

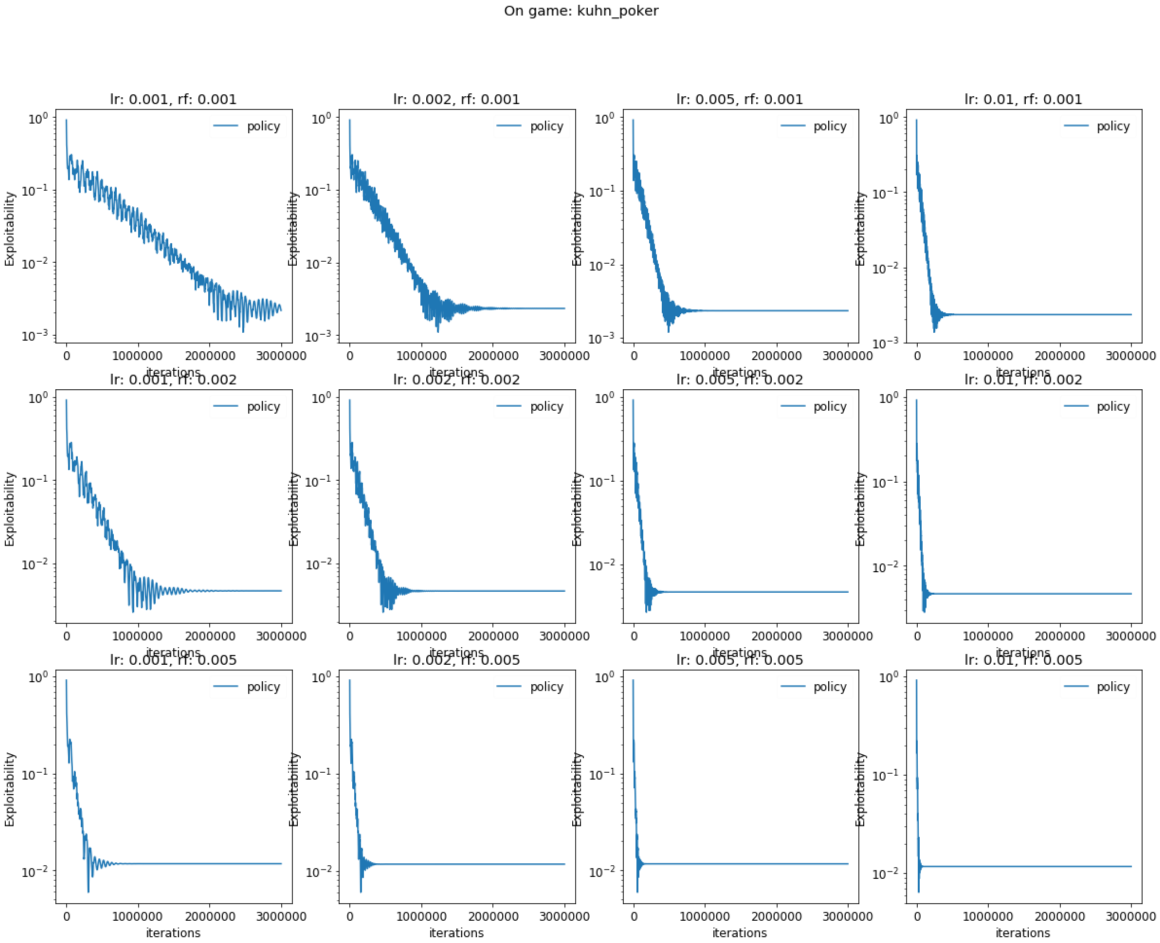

Figure 8. Kuhn Poker.

And the following experiment Shows the FoReL dynamics on Leduc poker :

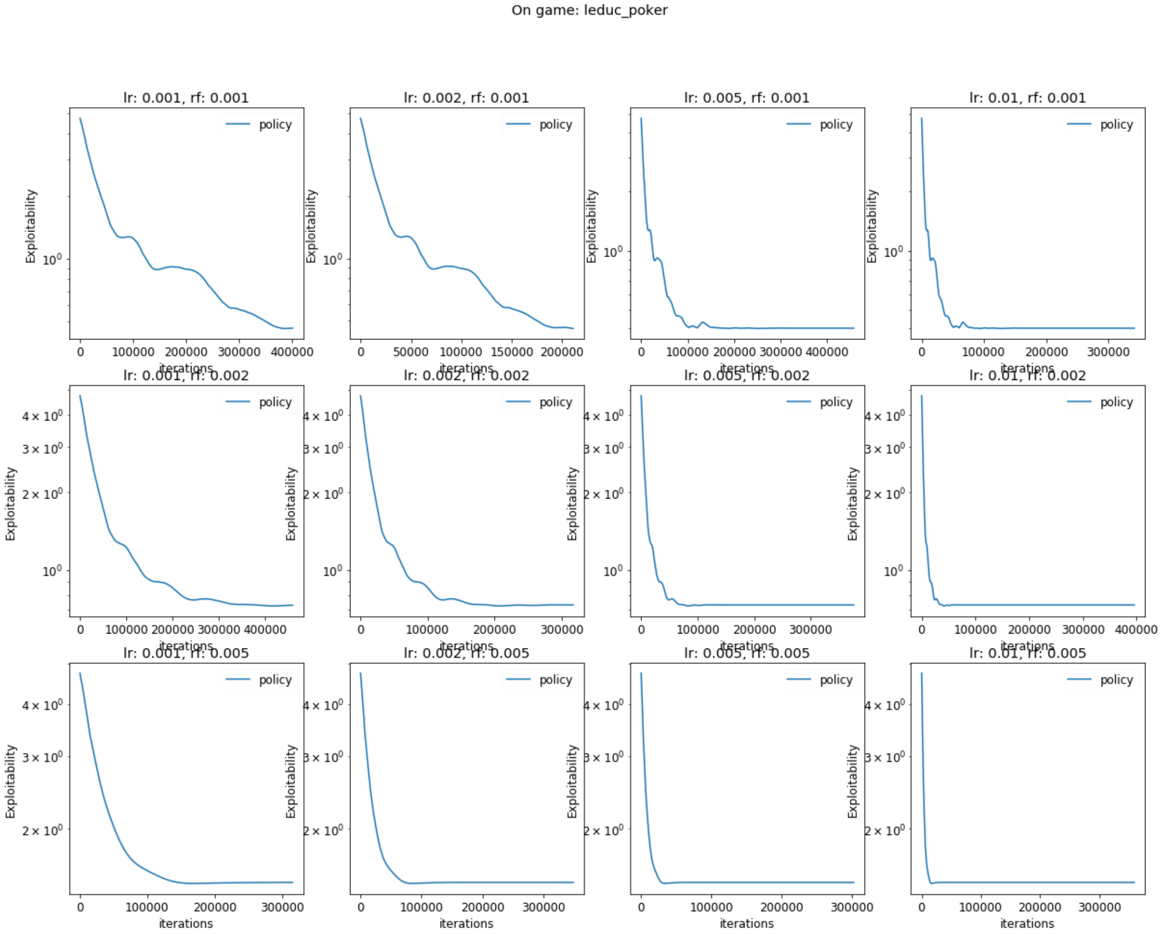

Figure 9. Leduc Poker.

K.1.3. Tabular Experiments With an addaptive Regularization (reward transformation for monotone games and the reward is changed every 20000 steps)

The two following figures illustrate the method described in section 6. And the following experiment Shows the FoReL dynamics on Kuhn poker :

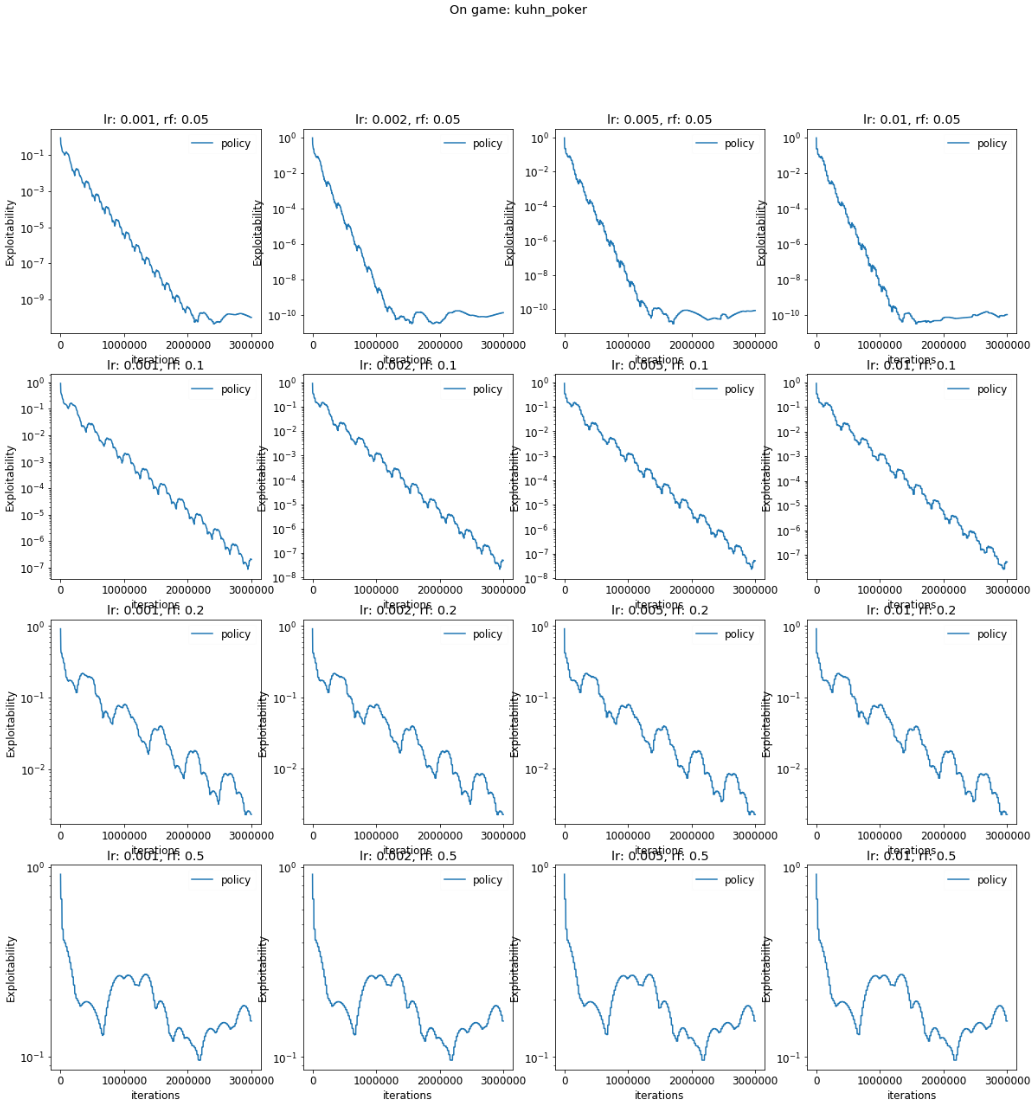

Figure 10. Kuhn Poker.

And the following experiment Shows the FoReL dynamics on Leduc poker :

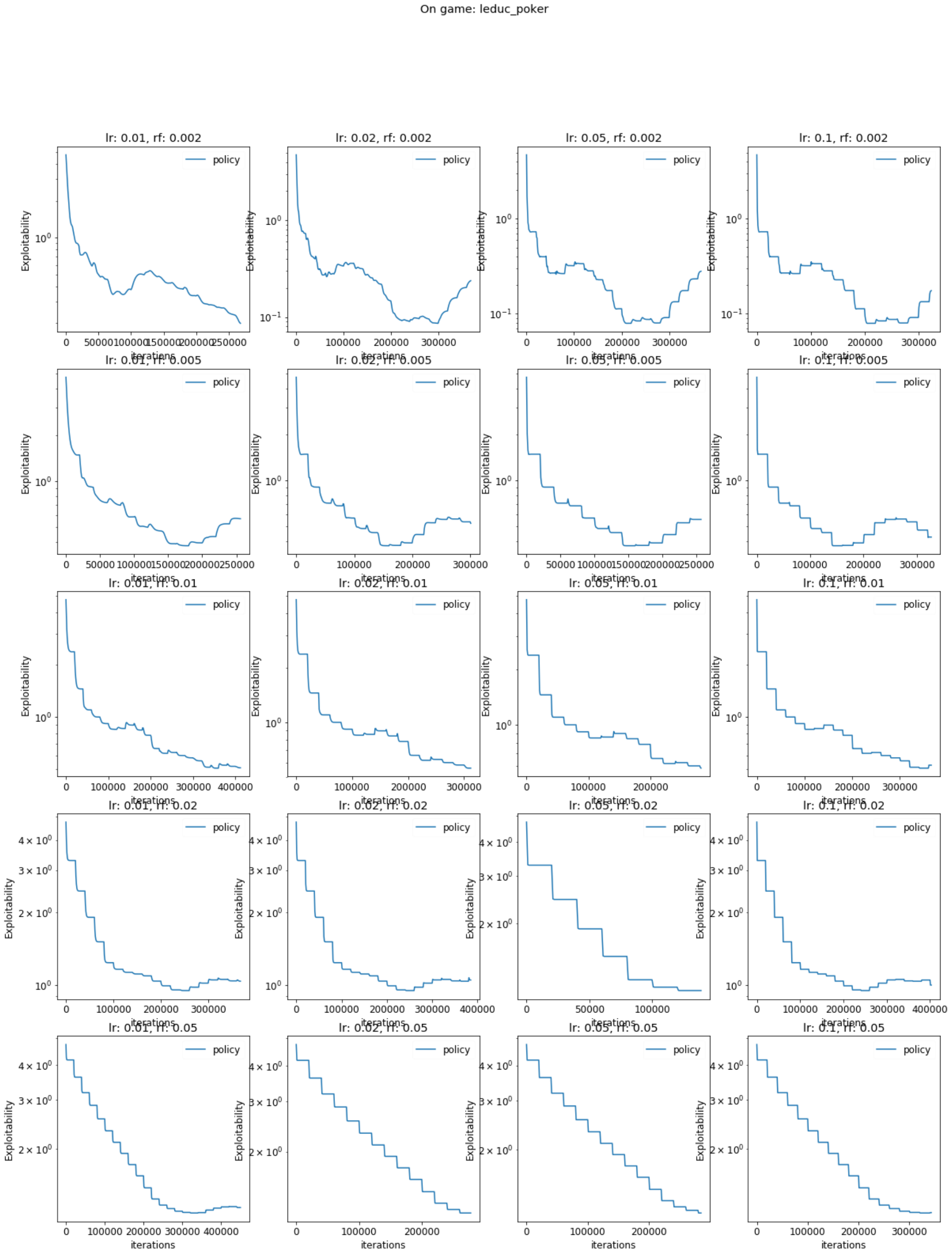

Figure 11. Leduc Poker.

K.2. Deep Reinforcement Learning Experiments with player only regularization

In this section, we run NeuRD on Leduc poker, Kuhn poker, Liars Dice and GoofSpiel with the reward transform for monotone games. The reward is adapted every 75000 steps.

batch size	256
batch size actor	32
eta	$\{1.0, 0.5, 0.2, 0.05, 0.02, 0.0\}$
lambda Retrace	1.0
epsilon greedy	0.1
Gradient clipping value	1000
max number of steps	4000000
reward re-centered every	75000
policy learning rate start	0.01
policy learning rate end	0.00001
threshold NeuRD	2
Neural net structure for π and Q	MLP with 2 hidden layer of 128 unit

Experiment on Liars Dice:

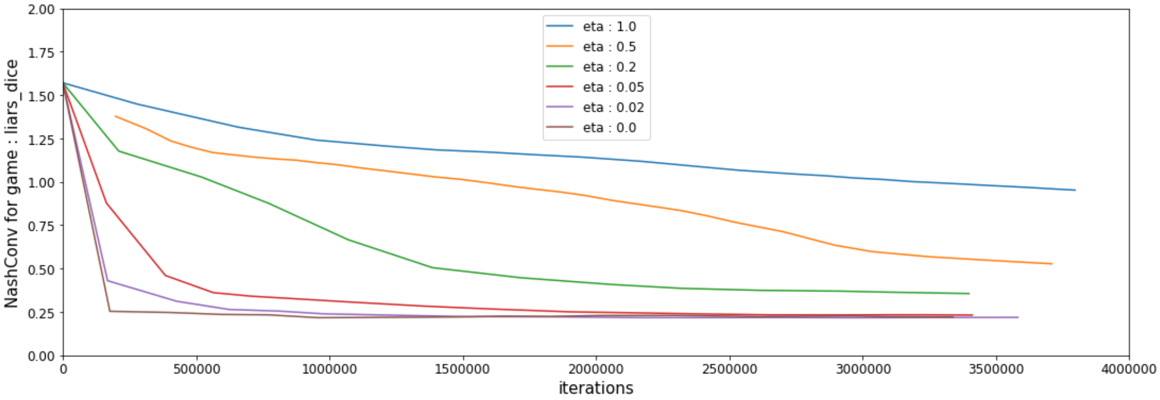

Figure 12. Liars Dice.

Experiment on Leduc Poker:

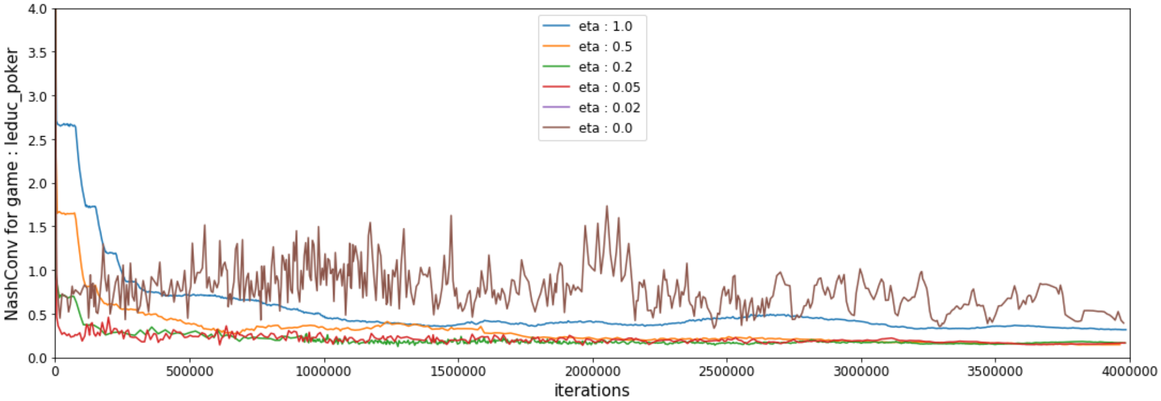

Figure 13. Leduc Poker.

Experiment on Kuhn Poker:

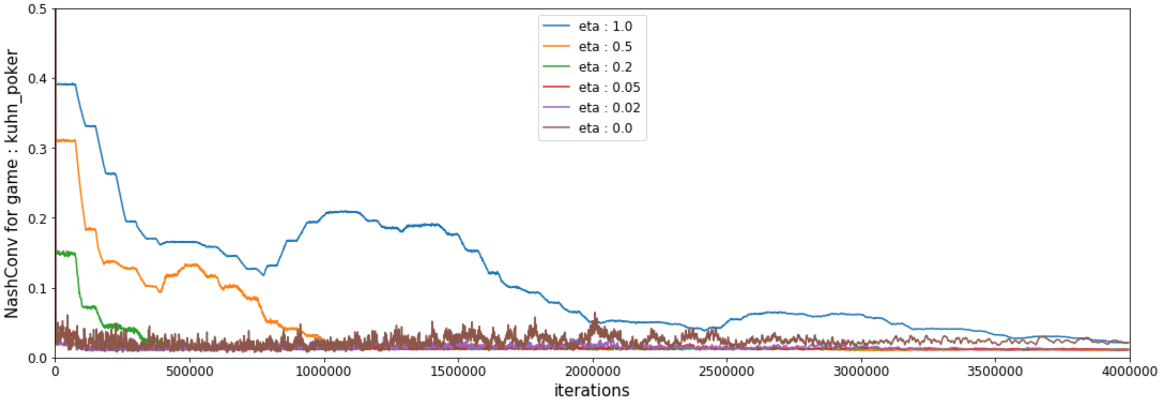

Figure 14. Kuhn Poker.

Experiment on Goofspiel:

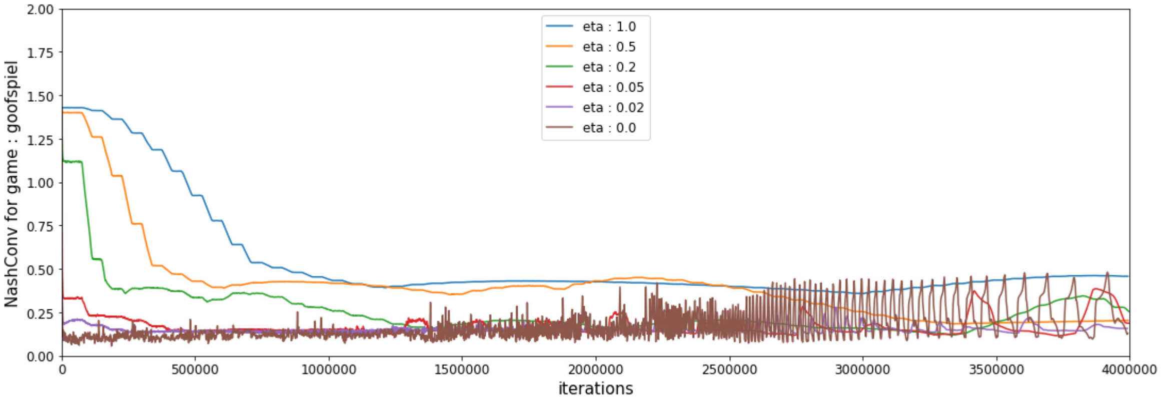

Figure 15. Goofspiel 4 cards.

In these experiments, we run NeuRD on Leduc poker, Kuhn poker, Liars Dice and GoofSpiel with the reward transform for monotone games with a constant regularization.

batch size	256
batch size actor	32
eta	$\{1.0, 0.5, 0.2, 0.05, 0.02, 0.0\}$
lambda Retrace	1.0
epsilon greedy	0.1
Gradient clipping value	1000
max number of steps	4000000
reward re-centered every	Never
policy learning rate start	0.01
policy learning rate end	0.00001
threshold NeuRD	2
Neural net structure for π and Q	MLP with 2 hidden layer of 128 unit

Experiment on Liars Dice:

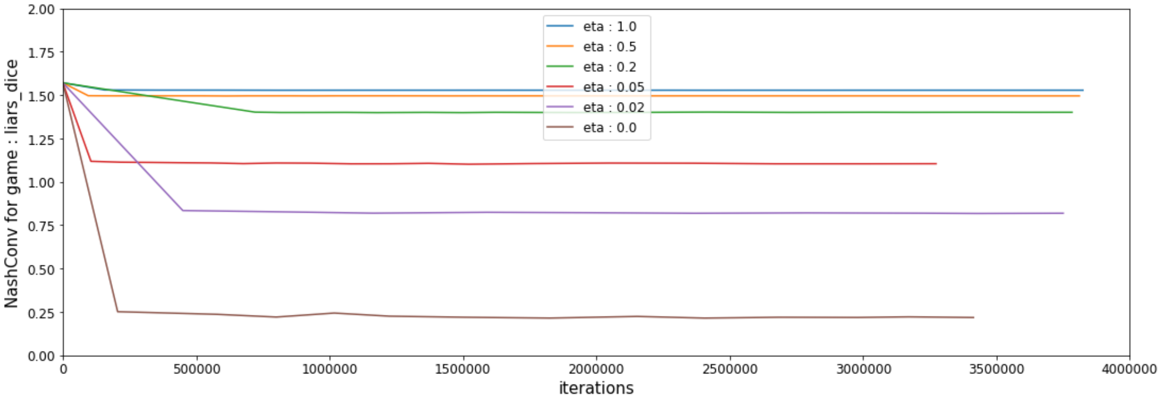

Figure 16. Liars Dice.

Experiment on Leduc Poker:

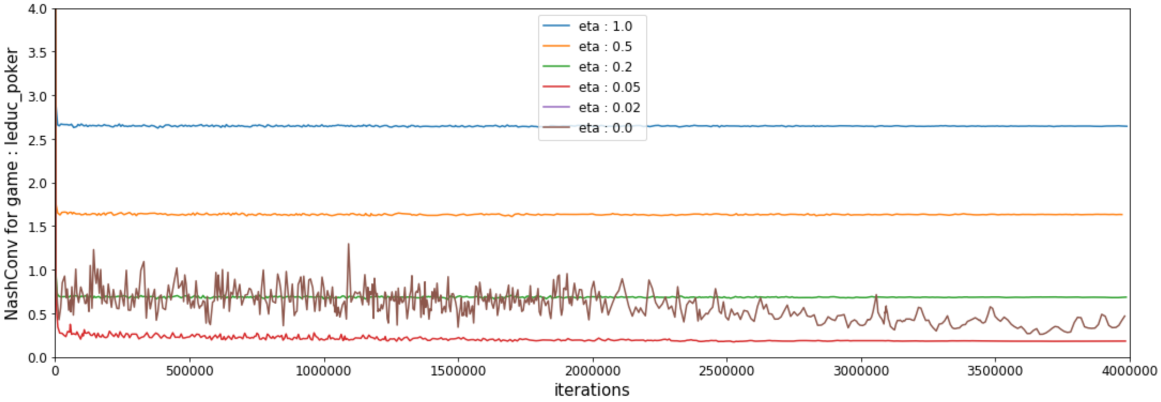

Figure 17. Leduc Poker.

Experiment on Kuhn Poker:

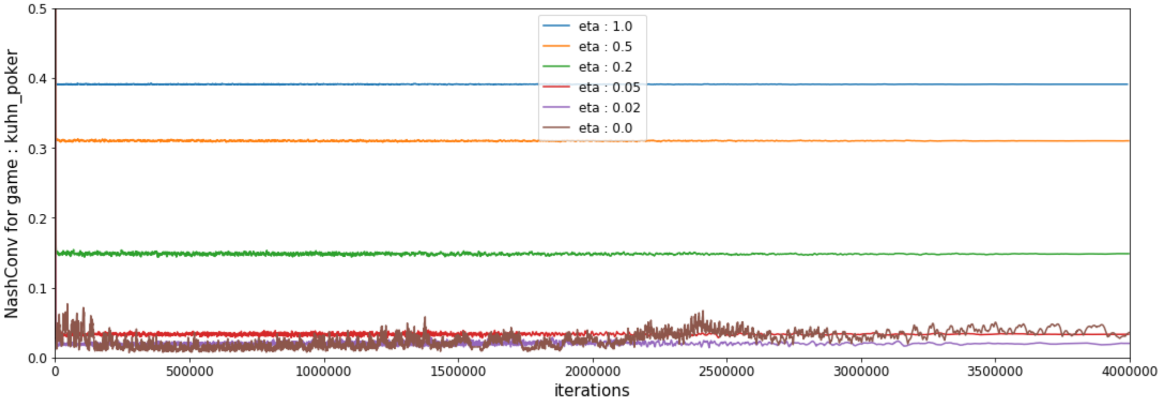

Figure 18. Kuhn Poker.

Experiment on Goofspiel:

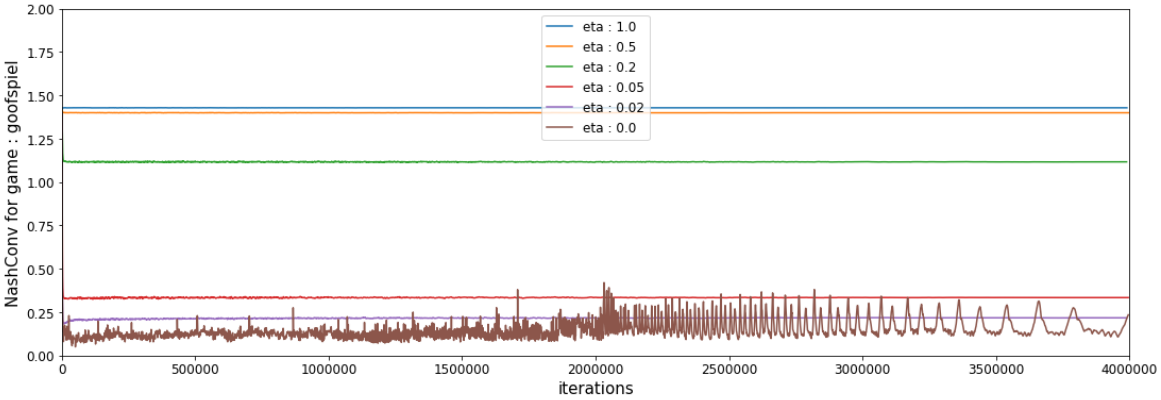

Figure 19. Goofspiel 4 cards.

In these experiments, we run NeuRD on Leduc poker, Kuhn poker, Liars Dice and GoofSpiel with the reward transform for monotone games with an exponential decay regularization to the regularization on the label.

batch size	256
batch size actor	32
eta exponential decay starting from 1.0 until the target value	$\{1.0, 0.5, 0.2, 0.05, 0.02, 0.0\}$
lambda Retrace	1.0
epsilon greedy	0.1
Gradient clipping value	1000
max number of steps	4000000
reward re-centered every	Never
policy learning rate start	0.01
policy learning rate end	0.00001
threshold NeuRD	2
Neural net structure for π and Q	MLP with 2 hidden layer of 128 unit

Experiment on Liars Dice:

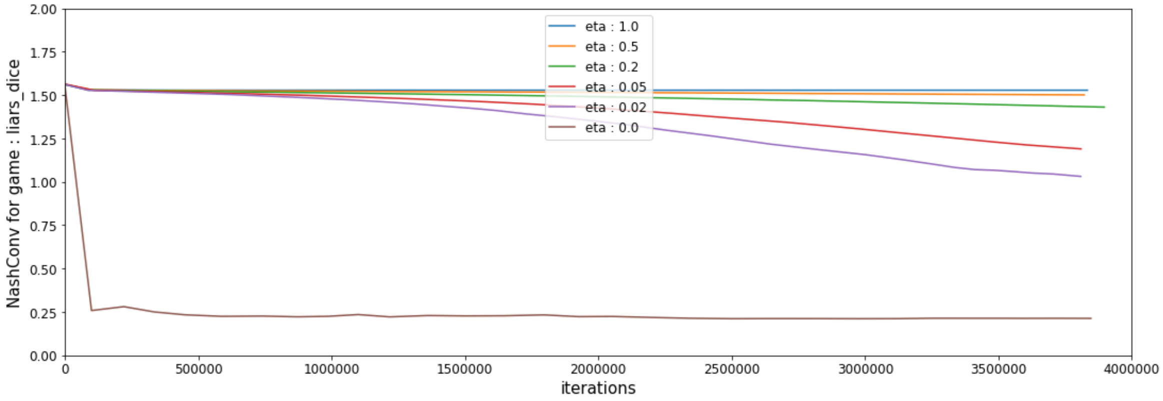

Figure 20. Liars Dice.

Experiment on Leduc Poker:

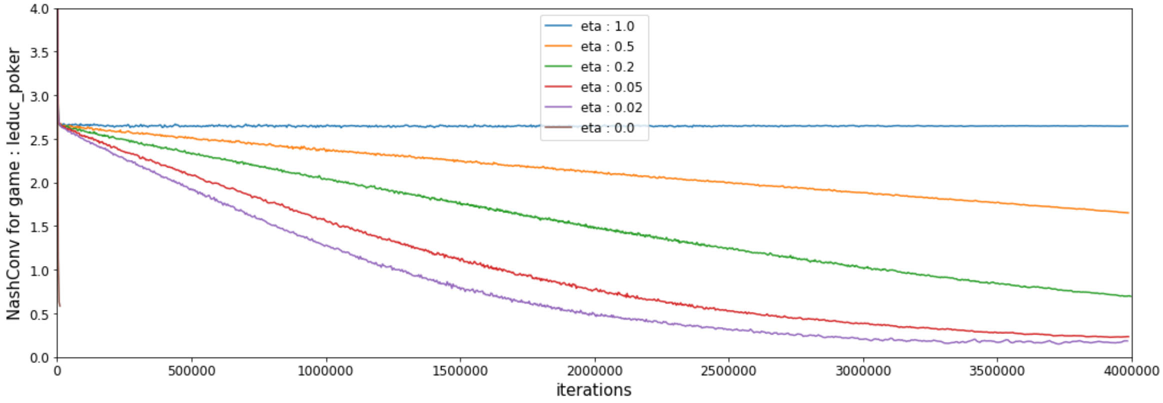

Figure 21. Leduc Poker.

Experiment on Kuhn Poker:

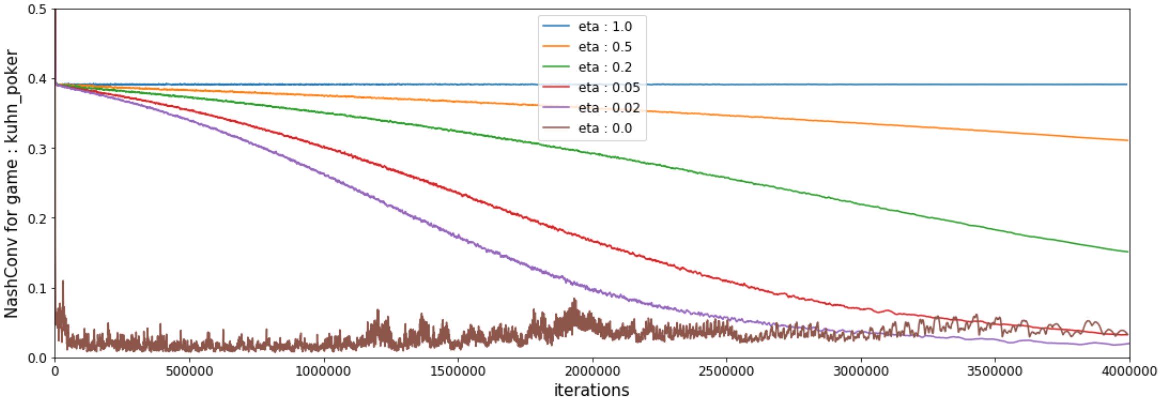

Figure 22. Kuhn Poker.

Experiment on Goofspiel:

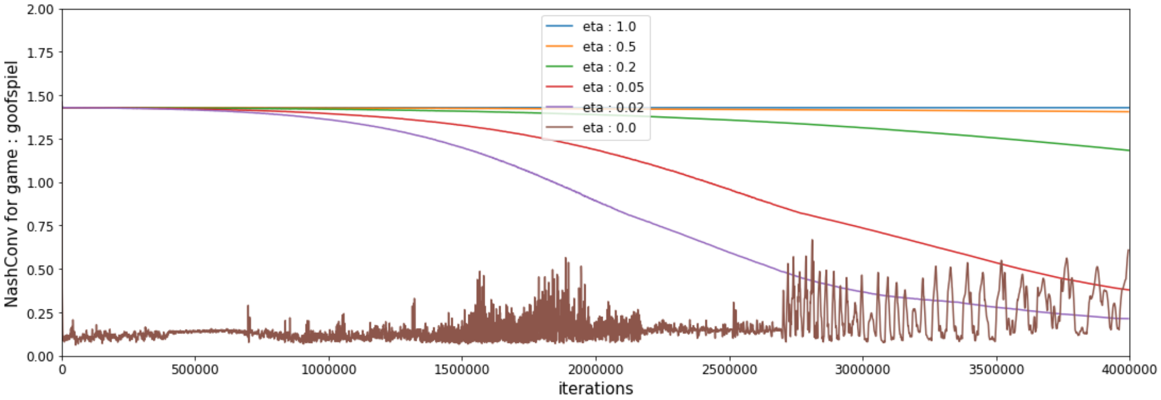

Figure 23. Goofspiel 4 cards.

K.3. Deep Reinforcement Learning Experiments with two player regularization

In these experiments, we run NeuRD on Leduc poker, Kuhn poker, Liars Dice and GoofSpiel with the reward transform for zero-sum games with an exponential decay regularization to the regularization on the label.

batch size	256
batch size actor	32
eta exponential decay starting from 1.0 until the target value	$\{1.0, 0.5, 0.2, 0.05, 0.02, 0.0\}$
lambda Retrace	1.0
epsilon greedy	0.1
Gradient clipping value	1000
max number of steps	4000000
reward re-centered every	Never
policy learning rate start	0.01
policy learning rate end	0.00001
threshold NeuRD	2
Neural net structure for π and Q	MLP with 2 hidden layer of 128 unit

Experiment on Liars Dice:

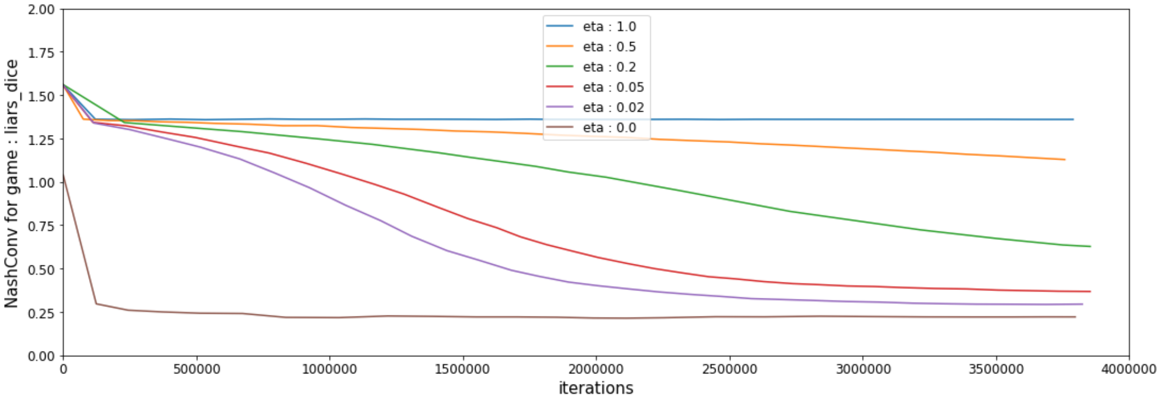

Figure 24. Liars Dice.

Experiment on Leduc Poker:

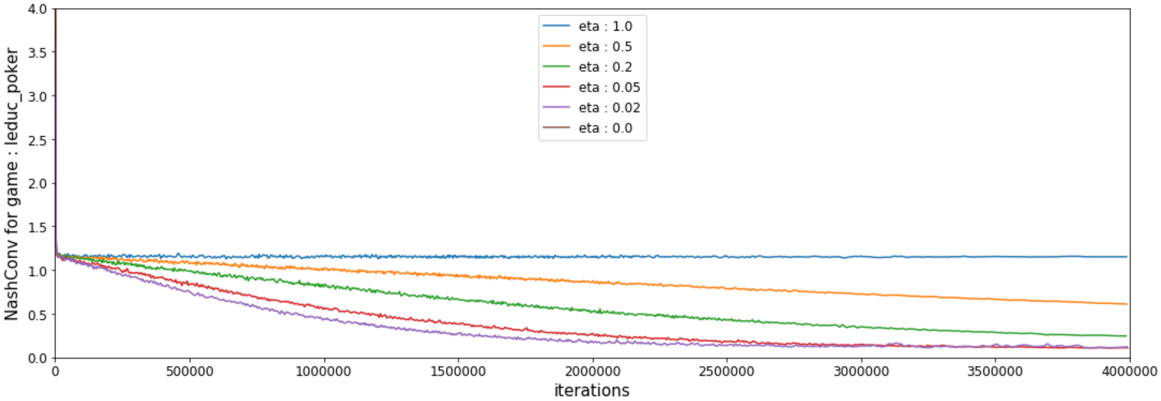

Figure 25. Leduc Poker.

Experiment on Kuhn Poker:

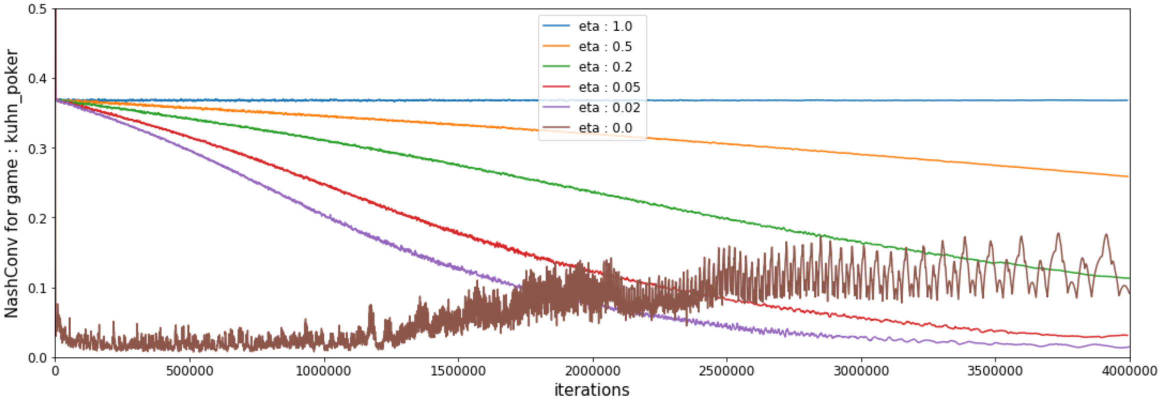

Figure 26. Kuhn Poker.

Experiment on Goofspiel:

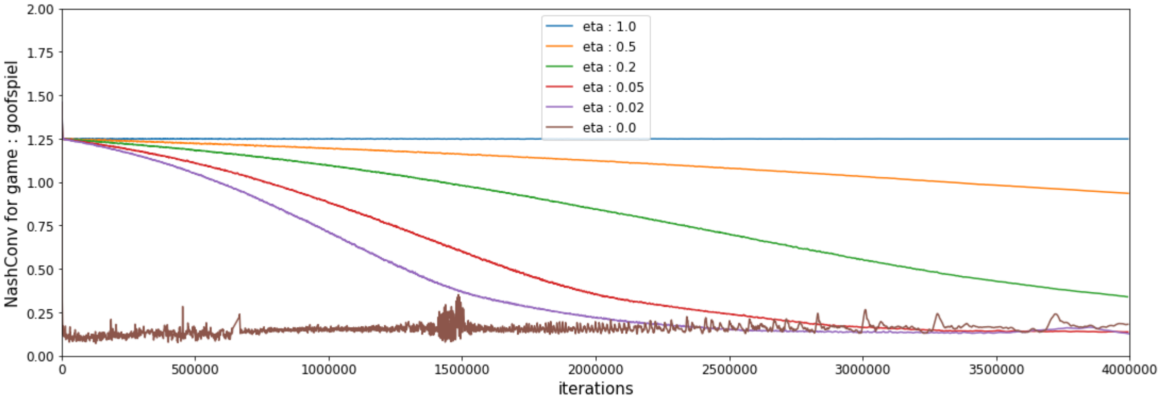

Figure 27. Goofspiel 4 cards.

K.4. Deep Reinforcement Learning Experiments with two player regularization with a large batch

In these experiments, we run NeuRD on Leduc poker, Kuhn poker, Liars Dice and GoofSpiel with the reward transform for zero-sum games with an exponential decay regularization to the regularization on the label.

batch size	2018
batch size actor	32
eta exponential decay starting from 1.0 until the target value	$\{1.0, 0.5, 0.2, 0.05, 0.02, 0.0\}$
lambda Retrace	1.0
epsilon greedy	0.1
Gradient clipping value	1000
max number of steps	4000000
reward re-centered every	Never
policy learning rate start	0.01
policy learning rate end	0.00001
threshold NeuRD	2
Neural net structure for π and Q	MLP with 2 hidden layer of 128 unit

Experiment on Liars Dice:

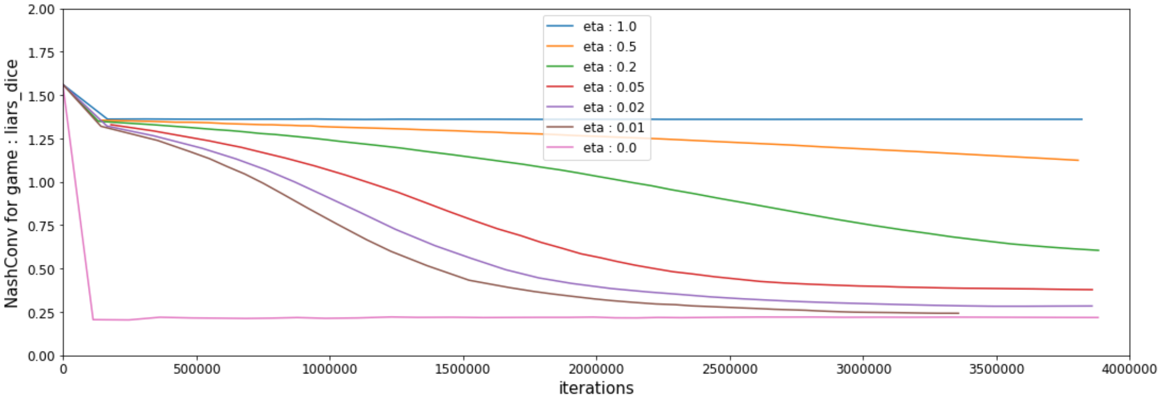

Figure 28. Liars Dice.

Experiment on Leduc Poker:

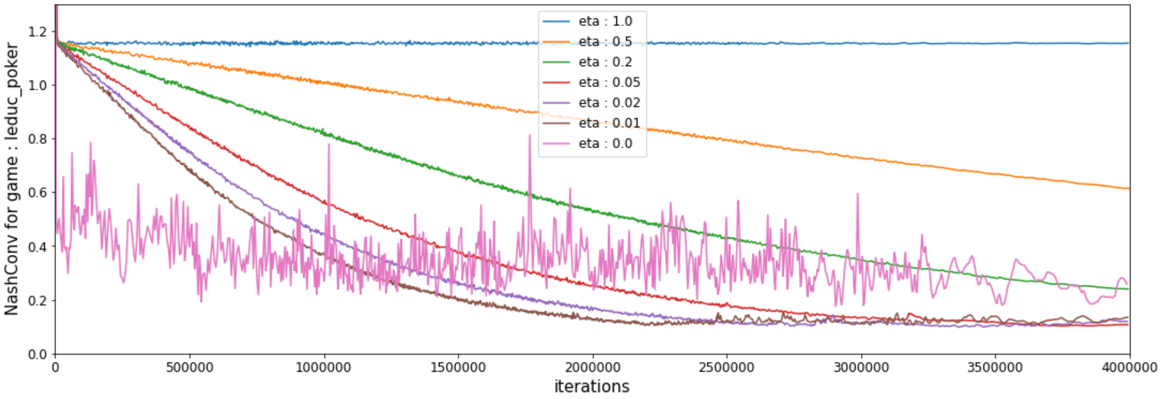

Figure 29. Leduc Poker.

Experiment on Kuhn Poker:

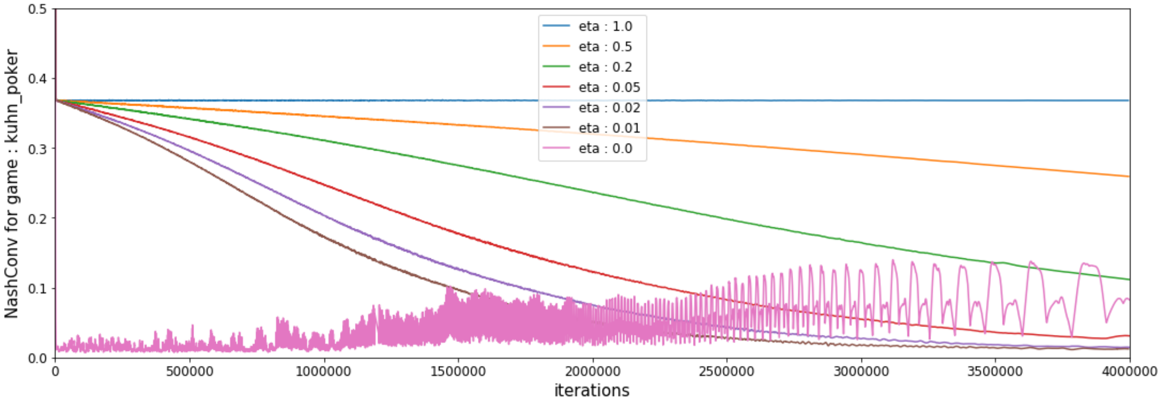

Figure 30. Kuhn Poker.

Experiment on Goofspiel:

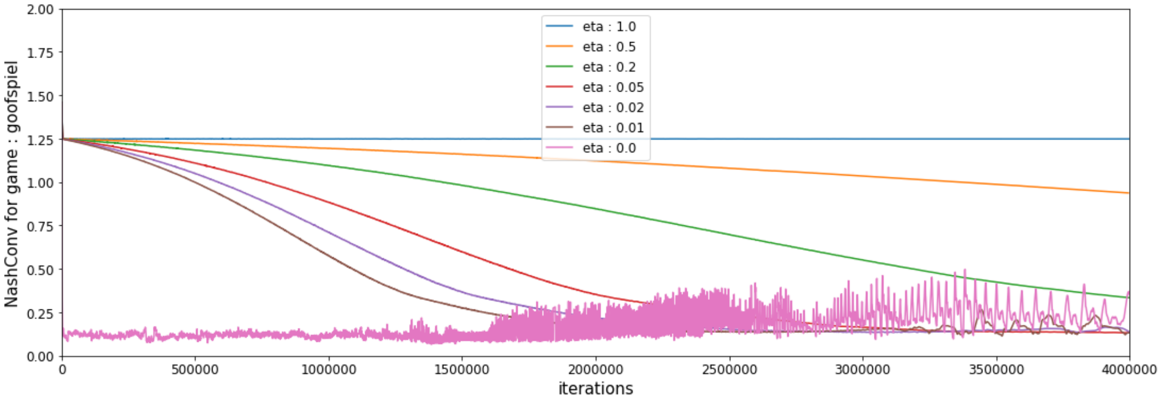

Figure 31. Goofspiel 4 cards.

L. Comparison to Baselines

We compared our method to 4 baselines implemented in OpenSpiel (Lanctot et al., 2019) with a comparable architecture (*i.e.* an MLP with two layer of 128 units)

L.1. DeepCFR

policy network layers	$[(128,), (128, 128)]$
advantage network layers	$[(128,), (128, 128)]$
num iterations	$[4000, 40000]$
num traversals	$[40, 400, 4000]$
learning rate	$[1e^{-3}]$
batch size advantage	$[128]$
batch size strategy	$[1024]$
memory capacity	$[1e7]$
policy network train steps	$[400]$
advantage network train steps	$[20, 40]$
reinitialize advantage networks	$[True, False]$

Figure 32. DeepCFR

L.2. NFSP

num train episodes	$[10e^6, 10e^7]$
eval every	$[10000]$
hidden layers sizes	$[[128], [128, 128]]$
replay buffer capacity	$[2e^5, 5e^5, 1e^6]$
reservoir buffer capacity	$[2e^6, 5e^6, 1e^7]$
anticipatory param	$[0.1, 0.05, 0.02, 0.01]$
epsilon start	$[0.06]$
epsilon end	$[0.001]$

Figure 33. NFSP

L.3. Q-learning

num train episodes	$[10e^6, 10e^7]$
eval every	$[10000]$
hidden layers sizes	$[[128], [128, 128]]$
replay buffer capacity	$[2e^5, 5e^5, 1e^6]$
batch size	$[64, 128, 256]$
epsilon start	$[0.06]$
epsilon end	$[0.001]$

Figure 34. Q-learning

L.4. PSRO

num train episodes	$[10e^6, 10e^7]$
meta strategy method	$[alpharank, uniform, nash, prd]$
number policies selected	$[1]$
sims per entry	$[1000]$
gpsro iterations	$[1000]$
symmetric game	$[False]$
rectifier	$['']$
training strategy selector	$[probabilistic, topkprobabilities, uniform]$
oracle type	$[DQN, PG]$
number training episodes	$[1e^4, 1e^5]$
self play proportion	$[0.0]$
hidden layer size	$[128]$
batch size	$[32]$
sigma	$[0.0]$
optimizer str	$[adam, sgd]$
loss str	$[qpg]$
num q before pi	$[8]$
n hidden layers	$[1, 2]$
entropy cost	$[0.001]$
critic learning rate	$[1e^{-2}]$
pi learning rate	$[1e^{-3}]$
dqn learning rate	$[1e^{-2}]$
update target network every	$[1000]$
learn every	$[10]$
seed	$[1]$

Figure 35. PSRO