Fictitious Play in Self Play

Alireza Kazemipour *

July 29, 2025

Abstract

Modelling human strategic behavior has been at the core of studies ranging from Computing Science [20], Economics [2] to Politics [7] and beyond. All of these seemingly diverse perspectives on prescribing a single explanation for how humans make decisions, have one certain property in common that is, although Expected Utility Theory [37] is a deficient model of describing humans strategic behaviors and models such as Prospect Theory [22] or, Cognitive Hierarchy [11] are better descriptions for humans strategic behaviors, they all agree on a single fact that the main drive for decision-making is *maximizing a sense of the utility* that the agents (for example humans) consider for the outcomes of events. The differences of these perspectives are based on what aspects they choose to investigate to justify the deviations that humans make from the ideal case of maximizing the expected utility [37]. On the other hand, the nature of the real-world events imply that almost everything happens more than one time (Infinite Monkey Theorem [5]) and if there are regularities among occurrence of an event in different timesteps then, the concept of *learning* could be leveraged to guarantee a certain kind of behavior through multiple encounters with a certain event [34]. In this project we try to study ***Fictitious Play*** (FP) as an approach of learning in repeated games and show that in contrast to the family of *no-regret* learning algorithms, this approach is more in line with the original *motive* of decision-making which is ***maximizing a sense of the utility***.

1 Motivation

The concept of learning is a very natural way of *prescribing* how agents should act and update their strategies based on the experiences gained do far in a repeated game [34]. One class of learning algorithms in repeated games is the family of *no-regret* algorithms [10]. Depending on the type of the regret that a no-regret learner chooses to optimize, different types of behaviors with different types of guarantees might be achieved (if at all). For example, [40] chose the counterfactual regret and showed that in a repeated game *with rational players*, minimizing this regret is equivalent to finding ϵ-Nash equilibria in zero-sum two player games. This method is widely accepted because at the end, the final performance is measured with respect to Nash equilibrium, meaning that since *rational agents* are expected utility maximizers [37], it does not matter how much the regret was minimized but how much the resulting performance is close to Nash equilibrium.

1.1 Counterfactual Regret Minimization vs Fictitious Play

If the closeness to Nash equilibrium is the ultimate goal of measuring the success of a learning algorithm for rational agents, then are no-regret learning algorithms the best to accomplish this goal? To answer this question, [13] took counterfactual regret minimization (CFR [40]) as a representative of no-regret learning algorithms and compared it against FP and empirically showed that as was promised, CFR gets closer to Nash equilibrium in zero-sum two-player games but, in any other type of games, FP resulted in a behavior closer to Nash equilibrium. They even empirically showed that there are zero-sum two-players games that FP gives even a better result than CFR! The summary of their results are shown in Figure 1.

*kazemipo@ualberta.ca

n	m	# games	# iterations	Avg. CFR ϵ	Avg. FP ϵ	Avg. difference in ϵ	Winner
2 (zs)	3	10,000	10,000	0.00139	0.00133	$5.945\times10^{-5} \pm 9.511\times10^{-6}$	FP
2 (zs)	5	10,000	10,000	0.00239	0.00261	$-2.219\times10^{-4} \pm 1.550\times10^{-5}$	CFR
2 (zs)	10	10,000	10,000	0.00282	0.00464	$-0.0018 \pm 2.277\times10^{-5}$	CFR
2	3	10,000	10,000	8.963×10^{-4}	8.447×10^{-4}	$5.155\times10^{-5} \pm 3.934\times10^{-5}$	FP
2	5	100,000	10,000	0.00383	0.00377	$6.000\times10^{-5} \pm 5.855\times10^{-5}$	FP
2	10	100,000	10,000	0.01249	0.01244	$4.865\times10^{-5} \pm 1.590\times10^{-4}$	Tie
3	3	100,000	10,000	0.00768	0.00749	$1.897\times10^{-4} \pm 1.218\times10^{-4}$	FP
3	5	100,000	10,000	0.02312	0.02244	$6.784\times10^{-4} \pm 2.454\times10^{-4}$	FP
3	10	10,000	10,000	0.05963	0.05574	0.0039 ± 0.0012	FP
4	3	100,000	10,000	0.01951	0.01950	$9.798\times10^{-6} \pm 2.195\times10^{-4}$	Tie
4	5	10,000	10,000	0.05121	0.04635	0.0049 ± 0.0011	FP
4	10	10,000	10,000	0.08315	0.06661	$0.0165 \pm 8.910\times10^{-4}$	FP
5	3	10,000	10,000	0.03505	0.03303	$0.0020 \pm 8.921\times10^{-4}$	FP
5	5	10,000	10,000	0.06631	0.05447	$0.0118 \pm 8.896\times10^{-4}$	FP
5	10	10,000	1,000	0.06350	0.04341	$0.0201 \pm 5.509\times10^{-4}$	FP

(a) CFR gets closer than FP to Nash equilibrium **only** in some zero-sum two-player games.

n	m	# games	# iterations	Avg. CFR ϵ	Avg. FP ϵ	Avg. difference in ϵ	Winner
2 (zs)	3	10,000	10,000	0.00139	0.00133	$5.945\times10^{-5} \pm 9.511\times10^{-6}$	FP
2 (zs)	5	10,000	10,000	0.00239	0.00261	$-2.219\times10^{-4} \pm 1.550\times10^{-5}$	CFR
2 (zs)	10	10,000	10,000	0.00282	0.00464	$-0.0018 \pm 2.277\times10^{-5}$	CFR
2	3	10,000	10,000	8.963×10^{-4}	8.447×10^{-4}	$5.155\times10^{-5} \pm 3.934\times10^{-5}$	FP
2	5	100,000	10,000	0.00383	0.00377	$6.000\times10^{-5} \pm 5.855\times10^{-5}$	FP
2	10	100,000	10,000	0.01249	0.01244	$4.865\times10^{-5} \pm 1.590\times10^{-4}$	Tie
3	3	100,000	10,000	0.00768	0.00749	$1.897\times10^{-4} \pm 1.218\times10^{-4}$	FP
3	5	100,000	10,000	0.02312	0.02244	$6.784\times10^{-4} \pm 2.454\times10^{-4}$	FP
3	10	10,000	10,000	0.05963	0.05574	0.0039 ± 0.0012	FP
4	3	100,000	10,000	0.01951	0.01950	$9.798\times10^{-6} \pm 2.195\times10^{-4}$	Tie
4	5	10,000	10,000	0.05121	0.04635	0.0049 ± 0.0011	FP
4	10	10,000	10,000	0.08315	0.06661	$0.0165 \pm 8.910\times10^{-4}$	FP
5	3	10,000	10,000	0.03505	0.03303	$0.0020 \pm 8.921\times10^{-4}$	FP
5	5	10,000	10,000	0.06631	0.05447	$0.0118 \pm 8.896\times10^{-4}$	FP
5	10	10,000	1,000	0.06350	0.04341	$0.0201 \pm 5.509\times10^{-4}$	FP

(b) There are zero-sum two player games and other type of games that FP is closer to Nash equilibrium than CFR.

Figure 1: CFR vs FP [13].

2 Fictitious Play

Fictitious Play was originally proposed as a way of computing Nash equilibria in zero-sum games [8, 32] and in this settings, players don't need to know the game they're playing nor the payoffs of others [19]. In addition to its elegance and simplicity, one nice feature of FP compared to no-regret algorithms is its directness toward computing the best response and Nash equilibrium as discussed in Section 1.

In FP, N players play a repeated game and in each round $t \in \mathbb{N}^+$, each agent $i, \forall i \in N$ plays a mixed strategy π_t^i such that:

$$\pi_{t+1}^i \in (1 - \alpha_{t+1})\pi_t^i + \alpha_{t+1} b^i(\pi_t^{-i}) \tag{1}$$

Where π_t^{-i} is the agent i's belief about the mixed the strategy profile of all other players at round t and *could be* the empirical distribution of their previous actions, $b^i(\pi_t^{-i})$ is the set of best responses of the player i to other players' mixed strategy π_t^{-i} that it had assumed for them in round t, $\alpha_t = \frac{1}{t}$ is the step size and π_0^{-i} is the initial belief about other players' strategies and π_0^i is the player's i initial mixed strategy.

2.1 Convergence of FP

The primary topic that should be discussed prior to performance when studying learning algorithms is if the algorithm is convergent at all or not. If FP is convergent, it converges to Nash equilibrium [32]. But when is it?

2.1.1 Shapley's Almost-Rock-Paper-Scissors

The biggest weakness of FP is its heavy reliance on the initial beliefs. For example, [33] introduced a non zero-sum variant of the game of Rock-Paper-Scissors shown in Figure 2. The unique Nash equilibrium of this game is for each player to play the mixed strategy $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ however, when π_0^1 is initialized to $(0, 0, \frac{1}{2})$ and π_0^2 is initialized to $(0, \frac{1}{2}, 0)$, it can be shown that the empirical play of this game never converges to any fixed distribution.

2.1.2 Provably Convergent FP

Despite of FP's sensitivity to initial beliefs, it is convergent in zero-sum games [32], Potential games[1] [23, 4], $2 \times n$ with generic payoffs games [3] and games that are solvable by iterated elimination of strictly dominated strategies [28].

[1] A game is said to be a potential game if the incentive of all players to change their strategy can be expressed using a single global function called the potential function [30].

	Rock	Paper	Scissors
Rock	0, 0	0, 1	1, 0
Paper	1, 0	0, 0	0, 1
Scissors	0, 1	1, 0	0, 0

Figure 2: Shapley's Almost-Rock-Paper-Scissors.

2.2 Approximation to FP

The original version of FP that was shown in Inclusion 1, as mentioned before, was introduced in [8, 32]. Later, [36] proved that the following version of FP is also convergent in games that original one is convergent if, $\epsilon_t \to 0$ as $t \to \infty$:

$$\pi^i_{t+1} \in (1 - \alpha_{t+1})\pi^i_t + \alpha_{t+1} b^i_{\epsilon_{t+1}}(\pi^{-i}_t) \tag{2}$$

Where the setting is exactly as before with the difference that $b^i_{\epsilon_{t+1}}(\pi^{-i}_t)$ is the set of ϵ_{t+1}-best responses of the player i to other players' mixed strategy π^{-i}_t that it had assumed for them in round t. "Intuitively, since these *mistakes* vanish asymptotically, such processes should also follow the best response in the limit" [26].

Also, [1] showed that the beliefs for the opponents' strategies should not be perfect and could be *perturbed* and still FP would be convergent in games that original FP was convergent:

$$\pi^i_{t+1} \in (1 - \alpha_{t+1})\pi^i_t + \alpha_{t+1} b^i(\pi^{-i}_t + M^i_{t+1}) \tag{3}$$

Where the setting is exactly as the original FP with the difference that M^i_{t+1} represents the perturbations to the beliefs of player i at round $t+1$ about other players' mixed strategy π^{-i}_t that it had assumed for them in round t with the additional condition that for all $T > 0$:

$$\lim_{t\to\infty} \sup_k \left\{ \left\| \sum_{i=t}^{k-1} \alpha_{i+1} M_{i+1} \right\| : \sum_{i=t}^{k-1} \alpha_i \leq T \right\} = 0$$

At last, [26] introduced the concept of **Generalized Weakened Fictitious Play** which says that not only Inclusion 2 and Inclusion 3 are convergent separately, but their combination shown in Inclusion 4, with the same set of conditions of each of the separately, is also convergent:

$$\pi^i_{t+1} \in (1 - \alpha_{t+1})\pi^i_t + \alpha_{t+1} b^i_{\epsilon_{t+1}}(\pi^{-i}_t + M^i_{t+1}) \tag{4}$$

3 FP in Extensive-Form Games (XFP)

All of the convergence proof of any variants of FP was established in normal-form games however, the practical scenarios happen in sequential format so, it is more appealing to have proof of convergence for FP in extensive-form games. One easy but completely infeasible and impractical solution could be to investigate the convergence of FP in the induced normal-form of the extensive-form representation however, the resulting exponential representation is far from any practical interest. In this section we try to show how the concept of ***realization equivalence*** that was initially introduced in [24] could be an elegant way of suppressing the need of transforming the extensive-form into the induced normal-form to prove the convergence of XFP and as a result, having the possibility to perform *behavioral strategies*!

3.1 Realization Equivalence

For any player $i \in N$ of a **prefect-recall** extensive-form game, each of their information sets $h^i \in I^i$ where I^i is the set of all of the information sets of the player i, uniquely defines a **sequence** σ_{h^i} of actions that the player is required to take in order to reach information set h^i. Let $\Sigma^i = \{\sigma_h : h \in I^i\}$ denote the set of such sequences for player i. Furthermore, let $\sigma_h a$ denote the sequence that extends σ_h with action a.

Definition 1 (Realization Plan [38])**.** A realization plan of player $i \in N$ is a function, $x : \Sigma^i \to [0, 1]$, such that $x(\varnothing) = 1$ and $\forall h \in I^i : x(\sigma_h) = \sum_{a \in \chi(h)} x(\sigma_h a)$. where χ is the action function.

For example a behavioral strategy π induces a realization plan $x_\pi(\sigma_h) = \prod_{(h',a) \in \sigma_h} \pi(h', a), \forall h \in I$, where the notation (h', a) disambiguates actions taken at different information sets. Similarly, a realization plan induces a behavioral strategy $\pi(h, a) = \frac{x(\sigma_h, a)}{x(\sigma_h)}$ where π is defined arbitrarily at information sets that are never visited, i.e. when $x(\sigma_h) = 0$.

The following definition and theorems connect an extensive-form game's behavioral strategies with mixed strategies of the equivalent normal-form representation:

Definition 2 (Realization Equivalence [17])**.** Two strategies π_1 and π_2 of a player are realization-equivalent if for any fixed strategy profile of the other players both strategies, π_1 and π_2, define the same probability distribution over the states of the game.

Theorem 1 ([38])**.** *Two strategies are realization-equivalent if and only if they have the same realization plan.*

Theorem 2 ([24])**.** *For a player with perfect recall, any mixed strategy is realization-equivalent to a behavioral strategy, and vice versa.*

3.2 XFP vs Normal-Form FP

In this section we derive a process in behavioral strategies that is realization equivalent to normal-form fictitious play and accomplishing the goal of remaining in the extensive-form space.

Lemma 1 ([17])**.** *Let π and β be two behavioral strategies, P and B two mixed strategies that are realization equivalent to π and β, x_κ the realization plan corresponding to the behavioral strategy κ, and $\gamma_1, \gamma_2 \in \mathbb{R}_{\geq 0}$ with $\gamma_1 + \gamma_2 = 1$. Then, for each information set $h \in I$,*

$$\mu(h) = \pi(h) + \frac{\gamma_2 x_\beta(\sigma_h)}{\gamma_1 x_\pi(\sigma_h) + \gamma_2 x_\beta(\sigma_h)}\big(\beta(h) - \pi(h)\big)$$

defines a behavioral strategy μ at h and μ is realization equivalent to the mixed strategy $M = \gamma_1 P + \gamma_2 B$.

Proof. The realization plan of $M = \gamma_1 P + \gamma_2 B$ is

$$x_M(\sigma_h) = \gamma_1 x_P(\sigma_h) + \gamma_2 x_B(\sigma_h), \forall h \in I$$

and due to realization equivalence, $x_P(\sigma_h) = x_\pi(\sigma_h)$ and $x_B(\sigma_h) = x_\beta(\sigma_h), \forall h \in I$. This realization plan induces a realization equivalent behavioral strategy:

$$\begin{aligned}\mu(h, a) &= \frac{x_M(\sigma_h a)}{x_M(\sigma_h)} = \frac{\gamma_1 x_\pi(\sigma_h a) + \gamma_2 x_\beta(\sigma_h a)}{\gamma_1 x_\pi(\sigma_h) + \gamma_2 x_\beta(\sigma_h)} = \frac{\gamma_1 x_\pi(\sigma_h)\pi(h, a) + \gamma_2 x_\beta(\sigma_h)\beta(h, a)}{\gamma_1 x_\pi(\sigma_h) + \gamma_2 x_\beta(\sigma_h)} \\ &= \pi(h, a) + \frac{\gamma_2 x_\beta(\sigma_h)}{\gamma_1 x_\pi(\sigma_h) + \gamma_2 x_\beta(\sigma_h)}\big(\beta(h, a) - \pi(h, a)\big)\end{aligned}$$

□

Now it is obvious that if we replace π with π_t^i, β with $b^i(\pi_t^{-i})$ (or its ϵ version), γ_1 with $1 - \alpha_{t+1}$ and γ_2 with α_{t+1}, the realization equivalence from behavioral strategies in extensive-from to mixed strategies in normal-from is accomplished in FP. So, the following theorem presents a fictitious play in behavioral strategies that inherits the convergence results of generalized weakened fictitious play by the virtue of realization equivalence.

Theorem 3 ([17]). *Let π_0 be an initial behavioral strategy profile. The extensive-form process*

$$\beta^i_{t+1} \in b^i_{\epsilon_{t+1}}(\pi_t^{-i})$$

$$\pi^i_{t+1}(h) = \pi^i_t(h) + \frac{\alpha_{t+1} x_{\beta^i_{t+1}}(\sigma_h)}{(1-\alpha_{t+1}) x_{\pi^i_t}(\sigma_h) + \alpha_{t+1} x_{\beta^i_{t+1}}(\sigma_h)} \big(\beta^i_{t+1}(h) - \pi^i_t(h)\big)$$

for all player $i \in N$ and all their information sets $h \in I$ is realization equivalent to a generalized weakened fictitious play in the normal-form, with the same convergence conditions of generalized weakened fictitious play [26], and therefore the average strategy profile converges to a Nash equilibrium in games that generalized weakened fictitious play does.

Proof. By induction. Assume the behavioral strategy π_t and the mixed strategy P_t are realization equivalent and $\beta_{t+1} \in b_{\epsilon_{t+1}}(\pi_t)$ is and ϵ_{t+1}-best response to π_t. By Kuhn's Theorem [24], let B_{t+1} be any mixed strategy that is realization equivalent to β_{t+1}. Then, B_{t+1} is an ϵ_{t+1}-best response to P_t in the normal form. By Lemma 1, the update in behavioral policies, π_{t+1} is realization equivalent to the following update in mixed strategies

$$P_{t+1} = (1-\alpha_{t+1})P_t + \alpha_{t+1} B_{t+1}$$

and thus follows a generalized weakened fictitious play. □

4 Fictitious Self-Play (FPS)

XFP, as was introduced in Section 3, suffers from the curse of dimensionality; At each iteration, computation needs to be performed at all states of the game irrespective of their relevance. However, generalised weakened fictitious play shown in Equation 4 only requires approximate best responses and even allows some perturbations in the updates.

FSP replaces the two fictitious play operations, best response computation and average strategy updating, with machine learning algorithms. Approximate best responses are learned by reinforcement learning from playing against the opponents' average strategies. The average strategy updates can be formulated as a supervised learning task, where each player learns a transition model of the average responses.

4.1 Estimating the Best Response

Consider an extensive-form game and some strategy profile π. Then for each player $i \in N$ the strategy profile of their opponents, π^{-i} defines an MDP, $\mathcal{M}(\pi^{-i})$ [35, 16]. The MDP's dynamics are given by the rules of the extensive-form game, the chance function and the opponents' fixed strategy profile. The rewards are given by the game's payoff function. An ϵ-optimal policy of the MDP, $\mathcal{M}(\pi^{-i})$, therefore yields an ϵ-best response of player i to the strategy profile π^{-i}. Thus the iterative computation of approximate best responses can be formulated as a sequence of MDPs to solve approximately, e.g. by applying reinforcement learning to samples of experience from the respective MDPs.

While generalised weakened fictitious play allows ϵ_t-best responses at iteration t, it requires that ϵ_t vanishes asymptotically such that $\epsilon_t \to 0$ as $t \to \infty$ [26, 36]. Corollary 1 bounds the absolute amount by which reinforcement learning needs to improve the best response profile to achieve a monotonic decay of the optimality gap ϵ_t.

Corollary 1 ([17]). *Let Π denote the set of behavioral strategies in a **two-player zero-sum** extensive-form game with maximum payoff range $\bar{R} = max_{\pi\in\Pi} R(\pi) - min_{\pi\in\Pi} R(\pi)$. Consider a fictitious play process in this game. Let P_t be the average strategy profile at iteration t, B_{t+1} a profile of ϵ_{t+1}-best responses to P_t, and $P_{t+1} = (1-\alpha_{t+1})P_t + \alpha_{t+1}B_{t+1}$ the usual fictitious play update for some stepsize $\alpha_{t+1} \in (0,1)$. Then for each player i, B^i_{t+1} is an $[\epsilon_t + \alpha_{t+1}(\bar{R} - \epsilon_t)]$-best response to P_{t+1}.*

[17] used Fitted Q-Learning [12] as the reinforcement learning method to find the best response. It learns from data sets of sampled experience; at each iteration t, FSP samples episodes of the game from **self-play** where each agent adds its experience to its replay memory. Each episode is in the form of, $\mathcal{E} = \{(h_{t'}, a_{t'}, r_{t'}, h_{t'+1})\}_{0\le t'\le T}, T \in \mathbb{N}$, where $h_{t'}, a_{t'}, r_{t'+1}$ correspond to the current information set, the action taken, the resulting payoff at the timestep t' of the episode recorded in round t.

4.2 Estimating the Opponents' Mixed Strategy[2]

Consider the point of view of a particular player i with the set of available actions $\mathcal{A}$ and the set of mixed strategies $S = \Delta(\mathcal{A})$ who wants to learn a behavioral strategy π that is realization equivalent to mixed strategy P which is a convex combination of their own normal-form mixed strategies, $P = \sum_{k=1}^{|S|} w_k.S(k)$ with $\sum_{k=1}^{|S|} w_k = 1$. This task is equivalent to learning a model of the player's behavior when it is sampled from P. Lemma 1 describes the behavioral strategy π explicitly, while in a sample-based setting we use samples from the realization equivalent strategy P to learn an approximation of π. We can sample from P by sampling from each constituent $S(k)$ with probability w_k.

Let $\tilde{\pi}_t^{-i}$ be the approximated behavioral strategy of the opponents' of player i at **round $\mathbf{t+1}$** (note that the best response of player i at round $t+1$ is with respect to its belief about opponents' round t strategy[3]) and $\tilde{P}_t^{-i}$ be its normal-form mixed strategy equivalent then, $\tilde{\pi}_t^{-i}$ is realization-equivalent to a perturbed fictitious play update in normal-form, $\pi_t^{-i} + \alpha_{t+1} M_{t+1}^i$ where $M_{t+1}^i = \frac{1}{\alpha_{t+1}}(\tilde{\pi}_t^{-i} - \pi_t^{-i})$ is a normal-form perturbation resulting from the estimation error.

[17] restrict themselves to simple models that count the number of times an action has been taken at an information state or alternatively accumulate the respective strategies' probabilities of taking each action. Their model update requires a set of sampled tuples, (h_t^i, ρ_t^i), where h_t^i is agent i's information set and ρ_t^i is the policy that the agent pursued at this information set when this experience was sampled. For each tuple (h_t^i, ρ_t^i) the update accumulates each action's weight at the information state,

$$\forall a \in \chi(h_t) : \mathcal{N}(h_t, a) \leftarrow \mathcal{N}(h_t, a) + \rho_t(a)$$
$$\forall a \in \chi(h_t) : \tilde{\pi}(h_t, a) \leftarrow \frac{\mathcal{N}(h_t, a)}{\mathcal{N}(h_t)}$$

The summary of SFP procedure is shown in Figure 3 and is results compared to XP are shown in Figure 4.

4.3 Neural SFP

By the virtue of approximation possibilities that SFP brought to life, more advanced estimation techniques can be applied to solve more complex and larg-scale problems using this paradigm. Neural Self Fictitious-Play (NSFP [18]) used Deep Reinforcement Learning [29, 31] and Neural Networks [25] for approximating the best responses and average strategies explained before in very large-scale problems thus, closer to real-world applications. For example, CFR [40], its extensions [6, 21] and other works [14] that tried to tackle the game of Limit Texas Hold'em Poker that has an order of 10^{17} states and 10^{14} information sets, hand-engineered the state representations and introduced hand-designed abstractions to reduce the state space size to be able to apply their methods. [39] used function approximation to abstract the game to a tractable size but, their full-width algorithm has to implicitly reason about all information sets at each iteration, which is prohibitively expensive in large domains. In contrast to all of those methods, NFSP focuses on the sample-based reinforcement learning setting where the game's states need not be exhaustively enumerated and the learner may not even have a model of the game's dynamics thus, scaling end-to-end to learn approximate Nash equilibria without prior domain knowledge. Furthermore, developments in continuous-action reinforcement learning [27] could enable NFSP to be applied to continuous-action games, which game-theoretic methods deal with a great level of sophistication [15].

Also later [9] introduced Deep CFR, a variant of CFR that deployed Neural Networks [25] to make CFR scalable to large problems and showed that their approach outperforms NSFP in terms of sample complexity, depicted in Figure 5b. However, (i) they compare their result in the game of Leduc Hold'em Poker which is a smaller variant of Limit Texas Hold'em Poker that NSFP demonstrated superhuman performance in, so it is not clear how their algorithm would be ranked in that larger game against NSFP and (ii) they report that:

[2] In the original work, the emphasis was on the agent learning **her own** average mixed strategy but, its rationale was basically estimating the opponent's strategy who is updating her own strategy in the same manner.

[3] Refer to Inclusion 4 for clarification.

Algorithm 2 General Fictitious Self-Play

function FICTITIOUSSELFPLAY(Γ, n, m)
  Initialize completely mixed π_1
  $\beta_2 \leftarrow \pi_1$
  $j \leftarrow 2$
  while within computational budget **do**
    $\eta_j \leftarrow$ MIXINGPARAMETER(j)
    $\mathcal{D} \leftarrow$ GENERATEDATA($\pi_{j-1}, \beta_j, n, m, \eta_j$)
    for each player $i \in \mathcal{N}$ **do**
      $\mathcal{M}^i_{RL} \leftarrow$ UPDATERLMEMORY($\mathcal{M}^i_{RL}, \mathcal{D}^i$)
      $\mathcal{M}^i_{SL} \leftarrow$ UPDATESLMEMORY($\mathcal{M}^i_{SL}, \mathcal{D}^i$)
      $\beta^i_{j+1} \leftarrow$ REINFORCEMENTLEARNING($\mathcal{M}^i_{RL}$)
      $\pi^i_j \leftarrow$ SUPERVISEDLEARNING($\mathcal{M}^i_{SL}$)
    end for
    $j \leftarrow j + 1$
  end while
  return π_{j-1}
end function

function GENERATEDATA(π, β, n, m, η)
  $\sigma \leftarrow (1 - \eta)\pi + \eta\beta$
  $\mathcal{D} \leftarrow n$ episodes $\{t_k\}_{1 \le k \le n}$, sampled from strategy profile σ
  for each player $i \in \mathcal{N}$ **do**
    $\mathcal{D}^i \leftarrow m$ episodes $\{t^i_k\}_{1 \le k \le m}$, sampled from strategy profile (β^i, σ^{-i})
    $\mathcal{D}^i \leftarrow \mathcal{D}^i \cup \mathcal{D}$
  end for
  return $\{\mathcal{D}^k\}_{1 \le k \le N}$
end function

Algorithm 3 FSP with FQI and simple counting model

Instantiate functions FICTITIOUSSELFPLAY and GENERATEDATA as in algorithm 2

function UPDATERLMEMORY($\mathcal{M}^i_{RL}, \mathcal{D}^i$)
  $\mathcal{T} \leftarrow$ Extract from $\mathcal{D}^i$ episodes that consist of transitions $(u_t, a_t, r_{t+1}, u_{t+1})$ from player i's point of view.
  Add $\mathcal{T}$ to $\mathcal{M}^i_{RL}$, replacing oldest data if the memory is full.
  return $\mathcal{M}^i_{RL}$
end function

function UPDATESLMEMORY($\mathcal{M}^i_{SL}, \mathcal{D}^i$)
  $\mathcal{D}^i_\beta \leftarrow$ Extract all episodes from $\mathcal{D}^i$ where player i chose their approximate best response strategy.
  $\mathcal{B} \leftarrow$ Extract from $\mathcal{D}^i_\beta$ data that consist of pairs (u_t, μ_t), where μ_t is player i's strategy at information state u_t at the time of sampling the respective episode.
  return $\mathcal{B}$
end function

function REINFORCEMENTLEARNING($\mathcal{M}^i_{RL}$)
  Initialize FQI with previous iteration's Q-values.
  $\beta \leftarrow$ FQI($\mathcal{M}^i_{RL}$)
  return β
end function

function SUPERVISEDLEARNING($\mathcal{M}^i_{SL}$)
  Initialize counting model from previous iteration.
  for each (u_t, μ_t) in $\mathcal{M}^i_{SL}$ **do**
    $\forall a \in \mathcal{A}(u_t) : N(u_t, a) \leftarrow N(u_t, a) + \mu_t(a)$
    $\forall a \in \mathcal{A}(u_t) : \pi(u_t, a) \leftarrow \frac{N(u_t, a)}{N(u_t)}$
  end for
  return π
end function

Figure 3: Summary of SFP [17].

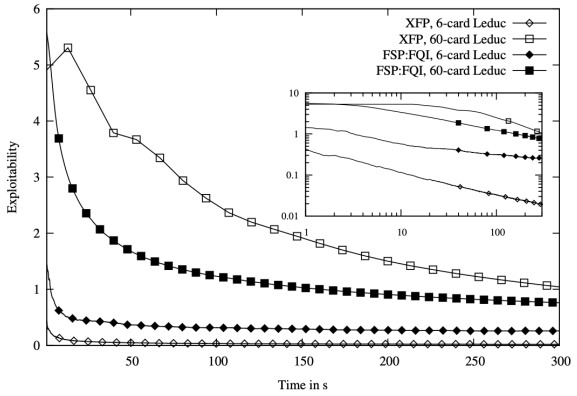
"We observe that Deep CFR reaches an exploitability of 37 mbb/g[4] while NFSP converges to 47 mbb/g. We also observe that Deep CFR is more sample efficient than NFSP. **However, these methods spend most of their wallclock time performing SGD steps, so in our implementation we see a less dramatic improvement over NFSP in wallclock time than sample efficiency**."

Figure 5 summarizes the performance of NSFP. [18] compared the performance of NSFP against SmooCT, one of the top 3 computer programs of Annual Computer Poker Competition that featured Limit Texas Hold'em Poker in 2014. [9] ran NFSP with the same model architecture as they used for Deep CFR to benchmark their sample complexity.

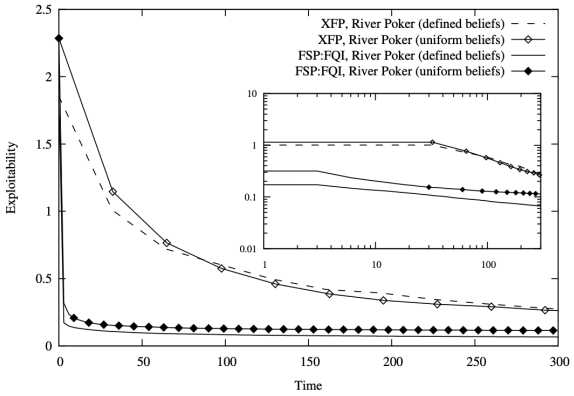
5 Conclusion

In this work, we tried to reiterate the correct objective when describing strategic behaviors using game theoretic perspectives is employed. The concept of *learning* is an inextricable tool in **describing** the pattern in the observed behavior or, **prescribing** how the desired behavior could be achieved over the course of time [34]. Thus, we reemphasized that learning should be studied with compatible motives of game theory, *maximizing a sense of the utility*, that for rational agents happens to be the expected utility. In this regard, we showed that no-regret algorithms, though achieving the desired goal, require delicate design and interpretations to relate to the desired maximization. As a replacement for this family of learning algorithms, we chose Fictitious play that because of incorporating the notion of *best*

[4] milli big blinds per game

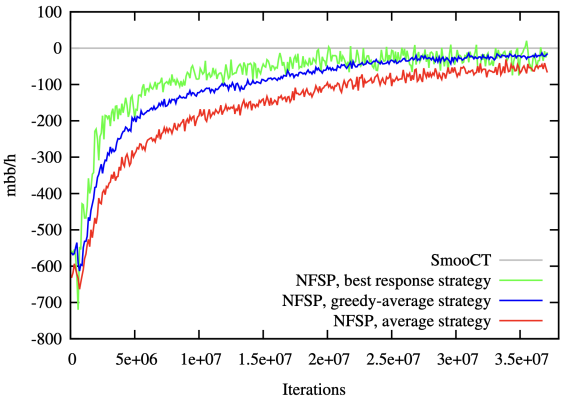


(a) Leduc Holdem. In small Games XPF is converging faster while in large games SFP.

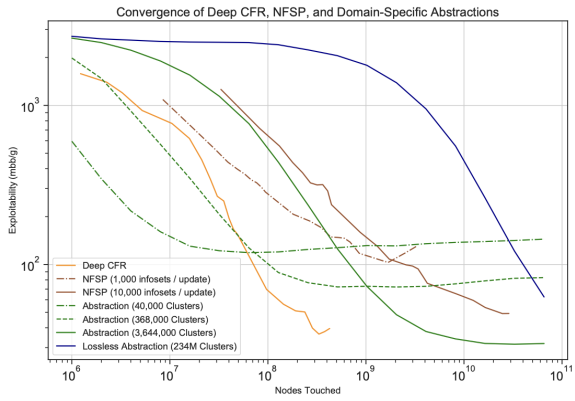


(b) River poker. The convergence rate of SFP is not sensitive to the initial beliefs.

Figure 4: Comparison of XFP and FSP-FQI. The inset presents the results using a logarithmic scale.



(a) Limit Texas Hold'em Poker. Win rates against SmooCT.



(b) Leduc Hold'em. Deep CFR achieves lower exploitability than NFSP with less samples.

Figure 5: NSFP performance.

response in its mechanism, is more in lined with game-theoretic perspectives. Fictitious Play allows to leverage approximations which in return opens the door to the world of many advanced and powerful machine learning algorithms useful for solving large-scale real-world problems that were not possible to do in classic game theory.

At last, we mentioned the state-of-the-art no-regret learning method that is also applicable to large-scale problems and we showed that although that method has the edge on its Fictitious Play counterpart in terms of sample complexity, their wallclock time as the quantity that is the main practical interest, are roughly the same. The no-regret algorithm was introduced almost 4 years later and it did not consider that NSFP could benefit from all of the advances happened in reinforcement learning literature in the meantime, and it used the original NSFP in their comparisons while as an advantage, Fictitious Play's compatibility with reinforcement learning makes this approach sustainable and naturally improved whenever a breakthrough occurs in the reinforcement learning community.

References

[1] Benaïm, M., Hofbauer, J., and Sorin, S. (2005). Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348.

[2] Benartzi, S. and Thaler, R. H. (1995). Myopic Loss Aversion and the Equity Premium Puzzle*. *The*

Quarterly Journal of Economics, 110(1):73–92.

[3] Berger, U. (2005). Fictitious play in 2× n games. *Journal of Economic Theory*, 120(2):139–154.

[4] Berger, U. (2007). Brown's original fictitious play. *Journal of Economic Theory*, 135(1):572–578.

[5] Borel, É. (1913). La mécanique statique et l'irréversibilité. *J. Phys. Theor. Appl.*, 3(1):189–196.

[6] Bowling, M., Burch, N., Johanson, M., and Tammelin, O. (2015). Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149.

[7] Brams, S. J. (2011). *Game theory and politics*. Courier Corporation.

[8] Brown, G. W. (1951). Iterative solution of games by fictitious play. *Act. Anal. Prod Allocation*, 13(1):374.

[9] Brown, N., Lerer, A., Gross, S., and Sandholm, T. (2019). Deep counterfactual regret minimization. In *International conference on machine learning*, pages 793–802. PMLR.

[10] Brown, W., Schneider, J., and Vodrahalli, K. (2023). Equilibrium Separations with No-Regret Agents.

[11] Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898.

[12] Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6.

[13] Ganzfried, S. (2020). Fictitious play outperforms counterfactual regret minimization. *arXiv preprint arXiv:2001.11165*.

[14] Gilpin, A., Hoda, S., Pena, J., and Sandholm, T. (2007). Gradient-based algorithms for finding nash equilibria in extensive form games. In *Internet and Network Economics: Third International Workshop, WINE 2007, San Diego, CA, USA, December 12-14, 2007. Proceedings 3*, pages 57–69. Springer.

[15] Glicksberg, I. L. (1952). A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points. *Proceedings of the American Mathematical Society*, 3(1):170–174.

[16] Greenwald, A., Li, J., Sodomka, E., and Littman, M. (2013). Solving for best responses in extensive-form games using reinforcement learning methods. In *Proceedings of the conference on reinforcement learning and decision making*, pages 116–120. Citeseer.

[17] Heinrich, J., Lanctot, M., and Silver, D. (2015). Fictitious self-play in extensive-form games. In *International conference on machine learning*, pages 805–813. PMLR.

[18] Heinrich, J. and Silver, D. (2016). Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*.

[19] Hendon, E., Jacobsen, H. J., and Sloth, B. (1994). Fictitious play in Extensive Form Games. Technical report.

[20] Hong, J., Levine, S., and Dragan, A. (2024). Learning to influence human behavior with offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

[21] Johanson, M., Burch, N., Valenzano, R., and Bowling, M. (2013). Evaluating state-space abstractions in extensive-form games. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 271–278.

[22] Kahneman, D. and Tversky, A. (1979). Prospect theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291.

[23] Krishna, V. (1992). *Learning in games with strategic complementarities*. Harvard Business School.

[24] Kuhn, H. W. (1953). Extensive games and the problem of information. *Contributions to the Theory of Games*, 2(28):193–216.

[25] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[26] Leslie, D. S. and Collins, E. J. (2006). Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298.

[27] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

[28] Miyasawa, K. (1961). *On the convergence of the learning process in a* 2×2 *non-zero-sum two-person game*. Princeton University Princeton.

[29] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

[30] Monderer, D. and Shapley, L. S. (1996). Potential Games. *Games and Economic Behavior*, 14(1):124–143.

[31] Riedmiller, M. (2005). Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16*, pages 317–328. Springer.

[32] Robinson, J. J. (1951). An iterative method of solving a game. *Classics in Game Theory*.

[33] Shapley, L. S. et al. (1963). *Some topics in two-person games*. Rand Corporation.

[34] Shoham, Y. and Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

[35] Silver, D. and Veness, J. (2010). Monte-carlo planning in large POMDPs. *Advances in neural information processing systems*, 23.

[36] Van der Genugten, B. (2000). A weakened form of fictitious play in two-person zero-sum games. *International Game Theory Review*, 2(04):307–328.

[37] Von Neumann, J. and Morgenstern, O. (2007). Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press.

[38] Von Stengel, B. (1996). Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246.

[39] Waugh, K., Morrill, D., Bagnell, J., and Bowling, M. (2015). Solving games with functional regret estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

[40] Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. (2007). Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20.