Fictitious Play in Self Play

Alireza Kazemipour

CMPUT 654: Modelling Human Strategic Behaviour

April 9th, 2024

Motivation

- What's all the fuss about?

Motivation

- What's all the fuss about?
  - ◇ We want to maximize a measure of our utility.

Motivation

- What's all the fuss about?
  - We want to maximize a measure of our utility.
- Why should we care about *learning*?[1]

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

Motivation

- What's all the fuss about?
  - We want to maximize a measure of our utility.
- Why should we care about *learning*?[1]
  - Temporal nature of the domain

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

Motivation

- What's all the fuss about?
  - We want to maximize a measure of our utility.
- Why should we care about *learning*?[1]
  - Temporal nature of the domain
  - Regularity across time

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

Motivation

- What's all the fuss about?
  - We want to maximize a measure of our utility.
- Why should we care about *learning*?[1]
  - Temporal nature of the domain
  - Regularity across time
  - Using strategies based on experiences gained so far

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

Motivation

- What's all the fuss about?
  - ◇ We want to maximize a measure of our utility.
- Why should we care about *learning*?[1]
  - ◇ Temporal nature of the domain
  - ◇ Regularity across time
  - ◇ Using strategies based on experiences gained so far
- Okay, calm down! We already have *no-regret* learning algorithms. They go toe-to-toe with humans in Poker![2] *:))*

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

[2] Zinkevich et al., "Regret minimization in games with incomplete information"

- What's all the fuss about?
  - We want to maximize a measure of our utility.
- Why should we care about *learning*?[1]
  - Temporal nature of the domain
  - Regularity across time
  - Using strategies based on experiences gained so far
- Okay, calm down! We already have *no-regret* learning algorithms. They go toe-to-toe with humans in Poker![2] *:))*
  - WHAT!!! Didn't you just say you want to *maximize a measure of your utility*?!

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

[2] Zinkevich et al., "Regret minimization in games with incomplete information"

- What's all the fuss about?
  - We want to maximize a measure of our utility.
- Why should we care about *learning*?[1]
  - Temporal nature of the domain
  - Regularity across time
  - Using strategies based on experiences gained so far
- Okay, calm down! We already have *no-regret* learning algorithms. They go toe-to-toe with humans in Poker![2] *:))*
  - WHAT!!! Didn't you just say you want to *maximize a measure of your utility*?!
  - Yeah, because closeness of their result to Nash equilibrium is still the final goal.

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

[2] Zinkevich et al., "Regret minimization in games with incomplete information"


FINDING
NASH EQUILIBRIA
EXPECTED
UTILITY
MAXIMIZATION
REGRET
MINIMIZATION
imgflip.com

Motivation

- Is CFR[2] (as a representative of no-regret algorithms) the best way to get close to Nash equilibrium?

[2] Zinkevich et al., "Regret minimization in games with incomplete information"

Motivation

- Is CFR[2] (as a representative of no-regret algorithms) the best way to get close to Nash equilibrium?
- Is there other ways of learning to achieve this goal better?

[2] Zinkevich et al., "Regret minimization in games with incomplete information"

Motivation

Is CFR the silver bullet?

Is CFR the silver bullet?[3]

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

[3] Ganzfried, "Fictitious play outperforms counterfactual regret minimization"

Is CFR the silver bullet?[3]

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

[3]Ganzfried, "Fictitious play outperforms counterfactual regret minimization"

Fictitious Play (FP)

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"
[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]
- The game Γ is repeated game.

[4] Robinson, "An iterative method of solving a game"

[5] Brown, "Iterative solution of games by fictitious play"

[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]
- The game Γ is repeated game.
- There are N players.

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"
[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]
- The game Γ is repeated game.
- There are N players.
- $b^i(\pi^{-i})$ the set of best responses of the player $i, \forall i \in N$ to other players' mixed strategy π^{-i}.

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"
[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]
- The game Γ is repeated game.
- There are N players.
- $b^i(\pi^{-i})$ the set of best responses of the player $i, \forall i \in N$ to other players' mixed strategy π^{-i}.
- Every player i, $\forall i \in N$ plays a mixed strategy π^i:

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"
[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]
- The game Γ is repeated game.
- There are N players.
- $b^i(\pi^{-i})$ the set of best responses of the player $i, \forall i \in N$ to other players' mixed strategy π^{-i}.
- Every player i, $\forall i \in N$ plays a mixed strategy π^i:
  - ◇ $\pi^i_{t+1} \in (1 - \alpha_{t+1})\pi^i_t + \alpha_{t+1} b^i(\pi^{-i}_t)$

[4] Robinson, "An iterative method of solving a game"

[5] Brown, "Iterative solution of games by fictitious play"

[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]
- The game Γ is repeated game.
- There are N players.
- $b^i(\pi^{-i})$ the set of best responses of the player $i, \forall i \in N$ to other players' mixed strategy π^{-i}.
- Every player i, $\forall i \in N$ plays a mixed strategy π^i:
  - $\pi^i_{t+1} \in (1 - \alpha_{t+1})\pi^i_t + \alpha_{t+1} b^i(\pi^{-i}_t)$
  - π^{-i}_t could be the empirical distribution of the opponent's previous actions.

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"
[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]
- The game Γ is repeated game.
- There are N players.
- $b^i(\pi^{-i})$ the set of best responses of the player $i, \forall i \in N$ to other players' mixed strategy π^{-i}.
- Every player i, $\forall i \in N$ plays a mixed strategy π^i:
  - $\pi^i_{t+1} \in (1 - \alpha_{t+1})\pi^i_t + \alpha_{t+1} b^i(\pi^{-i}_t)$
  - π^{-i}_t could be the empirical distribution of the opponent's previous actions.
  - $\alpha_t = \frac{1}{t}, t \in \mathbb{N}^+$

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"
[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Fictitious Play (FP)

- Initially, to compute Nash equilibria in zero-sum games.[4,5]
- Players don't need to know the game they're playing nor the payoffs of others.[6]
- The game Γ is repeated game.
- There are N players.
- $b^i(\pi^{-i})$ the set of best responses of the player $i, \forall i \in N$ to other players' mixed strategy π^{-i}.
- Every player i, $\forall i \in N$ plays a mixed strategy π^i:
  - $\pi^i_{t+1} \in (1 - \alpha_{t+1})\pi^i_t + \alpha_{t+1} b^i(\pi^{-i}_t)$
  - π^{-i}_t could be the empirical distribution of the opponent's previous actions.
  - $\alpha_t = \frac{1}{t}, t \in \mathbb{N}^+$
  - π^i_0: Initial beliefs.

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"
[6] Hendon, Jacobsen, and Sloth, *Fictitious Play in Extensive Form Games*

Example (π_t^{-i})

	Cooperate	Defect
Cooperate	-1,-1	-5,0
Defect	0,-5	-3,-3

Example (π_t^{-i})

	Cooperate	Defect
Cooperate	-1,-1	-5,0
Defect	0,-5	-3,-3

In a repeated Prisoner's Dilemma game, if the opponent has played C, C, D, C, D in the first five games, before the sixth game he is assumed to be playing the mixed strategy (0.6, 0.4).[1]

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

Example (Matching Pennies)

	Heads	Tails
Heads	$1, -1$	$-1, 1$
Tails	$-1, 1$	$1, -1$

Example (Matching Pennies)

	Heads	Tails
Heads	$1, -1$	$-1, 1$
Tails	$-1, 1$	$1, -1$

Round	1's action	2's action	1's beliefs	2's beliefs
0			(1.5,2)	(2,1.5)
1	T	T	(1.5,3)	(2,2.5)
2	T	H	(2.5,3)	(2,3.5)
3	T	H	(3.5,3)	(2,4.5)
4	H	H	(4.5,3)	(3,4.5)
5	H	H	(5.5,3)	(4,4.5)
6	H	H	(6.5,3)	(5,4.5)
7	H	T	(6.5,4)	(6,4.5)
⋮	⋮	⋮	⋮	⋮

Example (Matching Pennies)

	Heads	Tails
Heads	$1, -1$	$-1, 1$
Tails	$-1, 1$	$1, -1$

Round	1's action	2's action	1's beliefs	2's beliefs
0			(1.5,2)	(2,1.5)
1	T	T	(1.5,3)	(2,2.5)
2	T	H	(2.5,3)	(2,3.5)
3	T	H	(3.5,3)	(2,4.5)
4	H	H	(4.5,3)	(3,4.5)
5	H	H	(5.5,3)	(4,4.5)
6	H	H	(6.5,3)	(5,4.5)
7	H	T	(6.5,4)	(6,4.5)
⋮	⋮	⋮	⋮	⋮

Example (Matching Pennies)

	Heads	Tails
Heads	1, −1	−1, 1
Tails	−1, 1	1, −1

Round	1's action	2's action	1's beliefs	2's beliefs
0			(1.5,2)	(2,1.5)
1	T	T	(1.5,3)	(2,2.5)
2	T	H	(2.5,3)	(2,3.5)
3	T	H	(3.5,3)	(2,4.5)
4	H	H	(4.5,3)	(3,4.5)
5	H	H	(5.5,3)	(4,4.5)
6	H	H	(6.5,3)	(5,4.5)
7	H	T	(6.5,4)	(6,4.5)
⋮	⋮	⋮	⋮	⋮

Example (Matching Pennies)

	Heads	Tails
Heads	$1, -1$	$-1, 1$
Tails	$-1, 1$	$1, -1$

Round	1's action	2's action	1's beliefs	2's beliefs
0			(1.5,2)	(2,1.5)
1	T	T	(1.5,3)	(2,2.5)
2	T	H	(2.5,3)	(2,3.5)
3	T	H	(3.5,3)	(2,4.5)
4	H	H	(4.5,3)	(3,4.5)
5	H	H	(5.5,3)	(4,4.5)
6	H	H	(6.5,3)	(5,4.5)
7	H	T	(6.5,4)	(6,4.5)
⋮	⋮	⋮	⋮	⋮

Example (Matching Pennies)

	Heads	Tails
Heads	$1, -1$	$-1, 1$
Tails	$-1, 1$	$1, -1$

Round	1's action	2's action	1's beliefs	2's beliefs
0			(1.5,2)	(2,1.5)
1	T	T	(1.5,3)	(2,2.5)
2	T	H	(2.5,3)	(2,3.5)
3	T	H	(3.5,3)	(2,4.5)
4	H	H	(4.5,3)	(3,4.5)
5	H	H	(5.5,3)	(4,4.5)
6	H	H	(6.5,3)	(5,4.5)
7	H	T	(6.5,4)	(6,4.5)
⋮	⋮	⋮	⋮	⋮

Example (Matching Pennies)

Each player ends up alternating back and forth between playing heads and tails. In fact, as the number of rounds tends to infinity, the empirical distribution of the play of each player will converge to (0.5, 0.5).[1]

[1] Shoham and Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*

Example (Shapley's Almost-Rock-Paper-Scissors)

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

Example (Shapley's Almost-Rock-Paper-Scissors)

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

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 $\pi_0^1 = (0, 0, 0.5)$ and $\pi_0^2 = (0, 0.5, 0)$. It can be shown that the empirical play of this game never converges to any fixed distribution.[7]

[7] Shapley et al., *Some topics in two-person games*

Convergence of FP

Convergence of FP

- Zero-sum games.[4]

[4] Robinson, "An iterative method of solving a game"

Convergence of FP

- Zero-sum games[4].
- Potential Games[8,9].

[4] Robinson, "An iterative method of solving a game"
[8] Krishna, *Learning in games with strategic complementarities*
[9] Berger, "Brown's original fictitious play"

Convergence of FP

- Zero-sum games[4].
- Potential Games[8,9].
- $2 \times n$ with generic payoffs games[10].

[4] Robinson, "An iterative method of solving a game"
[8] Krishna, *Learning in games with strategic complementarities*
[9] Berger, "Brown's original fictitious play"
[10] Berger, "Fictitious play in 2× n games"

Convergence of FP

- Zero-sum games[4].
- Potential Games[8,9].
- $2 \times n$ with generic payoffs games[10].
- Solvable by iterated elimination of strictly dominated strategies games[11]Miyasawa, *On the convergence of the learning process in a* 2×2 *non-zero-sum two-person game*

[4] Robinson, “An iterative method of solving a game”
[8] Krishna, *Learning in games with strategic complementarities*
[9] Berger, “Brown’s original fictitious play”
[10] Berger, “Fictitious play in 2× n games”
[11] .

Variants of FP

Variants of FP

- Original FP[4,5]:

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

[4] Robinson, "An iterative method of solving a game"
[5] Brown, "Iterative solution of games by fictitious play"

Variants of FP

- Original FP[4,5]:

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

- ϵ-best response[12]:

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

[4] Robinson, "An iterative method of solving a game"

[5] Brown, "Iterative solution of games by fictitious play"

[12] Van der Genugten, "A weakened form of fictitious play in two-person zero-sum games"

Variants of FP

- Original FP[4,5]:

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

- ϵ-best response[12]:

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

- Perturbed best response[13]:

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

---

[4] Robinson, "An iterative method of solving a game"

[5] Brown, "Iterative solution of games by fictitious play"

[12] Van der Genugten, "A weakened form of fictitious play in two-person zero-sum games"

[13] Benaïm, Hofbauer, and Sorin, "Stochastic approximations and differential inclusions"

Generalized Weakened Fictitious Play[14]:

$$\pi_{t+1}^i \in (1-\alpha_{t+1})\pi_t^i + \alpha_{t+1} b_{\epsilon_t}^i(\pi_t^{-i} + M_{t+1}^i)$$

[14] Leslie and Collins, "Generalized weakened fictitious play"

Recap

Original Fictitious Play

Recap

Original Fictitious Play $\longrightarrow$ ϵ-best response:

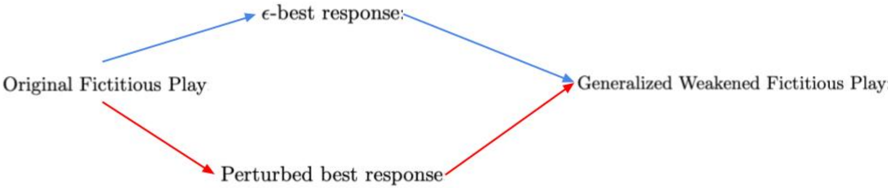
Recap

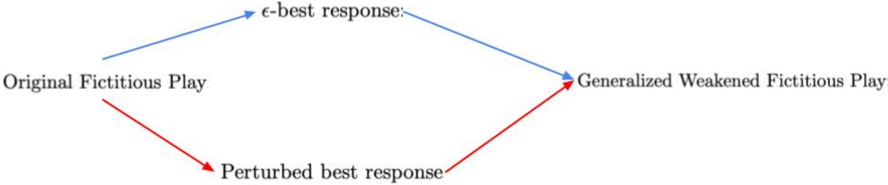
ϵ-best response:

Original Fictitious Play

Perturbed best response

Recap


ϵ-best response:
Original Fictitious Play
Generalized Weakened Fictitious Play
Perturbed best response


ϵ-best response:
Original Fictitious Play
Generalized Weakened Fictitious Play
Perturbed best response

WHY???

Let's shift gears!

- If FP in extensive-form (XPF) is *realization equivalent* to a normal-form FP then it inherits its convergence guarantees.

[15] Heinrich, Lanctot, and Silver, “Fictitious self-play in extensive-form games”

Convergence of FP in Extensive-form games

- If FP in extensive-form (XPF) is *realization equivalent* to a normal-form FP then it inherits its convergence guarantees.
- However, it can be implemented using only behavioral strategies and therefore its computational complexity per iteration is linear in the number of game states rather than exponential[15]! *:))*

[15] Heinrich, Lanctot, and Silver, "Fictitious self-play in extensive-form games"

Realization Equivalent

Realization Equivalent

Definition (Heinrich, Lanctot, and Silver, 2015)

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.

Realization Equivalent

Definition (Heinrich, Lanctot, and Silver, 2015)

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.

Definition (Kuhn, 1953)

For a player with perfect recall, any mixed strategy is realization-equivalent to a behavioral strategy, and vice versa.

Realization Equivalent

Lemma (Heinrich, Lanctot, and Silver, 2015)

Let π and β be two behavioral strategies, P and B two mixed strategies that realization equivalent to π and β, $\gamma_1, \gamma_2 \in \mathbb{R}_{\geq 0}$ with $\gamma_1 + \gamma_2 = 1$ and $x_\kappa(h)$ be the probability that a behavioral strategy κ get to an information set $h, \forall h \in I$ where I is the set of all information sets. Then $\forall h$:

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

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

XFP is realization equivalent to normal-form FP

Theorem (Heinrich, Lanctot, and Silver, 2015)

Let π_0 be an initial behavioral strategy profile. The extensive-form process:

$$\beta_t^i \in b_{\epsilon_t}^i(\pi_t^{-i})$$

$$\pi_{t+1}^i(h) = \pi_t^i(h) + \frac{\alpha_{t+1} x_{\beta_t}^i(h)}{(1-\alpha_{t+1}) x_{\pi_t}^i(h) + \alpha_{t+1} x_{\beta_t}^i(h)} (\beta_t^i(h) - \pi_t^i(h))$$

for all players $i \in N$ and all their information sets $h \in I^i$ is realization-equivalent to a generalized weakened fictitious play in the normal-form and therefore the average strategy profile converges to a Nash equilibrium.

XFP is realization equivalent to normal-form FP

Theorem (Heinrich, Lanctot, and Silver, 2015)

Let π_0 be an initial behavioral strategy profile. The extensive-form process:

$$\beta_t^i \in b_{\epsilon_t}^i(\pi_t^{-i})$$

$$\pi_{t+1}^i(h) = \pi_t^i(h) + \frac{\alpha_{t+1} x_{\beta_t}^i(h)}{(1-\alpha_{t+1}) x_{\pi_t}^i(h) + \alpha_{t+1} x_{\beta_t}^i(h)} (\beta_t^i(h) - \pi_t^i(h))$$

for all players $i \in N$ and all their information sets $h \in I^i$ is realization-equivalent to a generalized weakened fictitious play in the normal-form and therefore the average strategy profile converges to a Nash equilibrium.

Upshot: we can remain in the regime of behavioral strategies and apply FP! *:))*

In short

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

In short

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

$$\begin{cases} \beta^i_t \in b^i_{\epsilon_t}(\pi^{-i}_t) \\ \pi^i_{t+1}(h) = \pi^i_t(h) + \underbrace{\dfrac{\alpha_{t+1} x^i_{\beta_t}(h)}{(1-\alpha_{t+1})x^i_{\pi_t}(h) + \alpha_{t+1} x^i_{\beta_t}(h)}}_{\alpha_{t+1}} (\beta^i_t(h) - \pi^i_t(h)) \end{cases}$$

Fictitious Self-Play

XFP (like CFR[2]) sweeps the whole game tree. Can we make it more efficient?

[2] Zinkevich et al., "Regret minimization in games with incomplete information"

Generalized Weakened Fictitious Play[14]:

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

[14] Leslie and Collins, "Generalized weakened fictitious play"

Fictitious Self-Play

Generalized Weakened Fictitious Play[14]:

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

Generalized weakened fictitious play made leveraging two approximations possible that Fictitious Self-Play (FSP) implemented:

[14] Leslie and Collins, "Generalized weakened fictitious play"

Fictitious Self-Play

Generalized Weakened Fictitious Play[14]:

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

Generalized weakened fictitious play made leveraging two approximations possible that Fictitious Self-Play (FSP) implemented:

1. We can estimate the best response up to an ϵ_t error in round t.

[14] Leslie and Collins, "Generalized weakened fictitious play"

Fictitious Self-Play

Generalized Weakened Fictitious Play[14]:

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

Generalized weakened fictitious play made leveraging two approximations possible that Fictitious Self-Play (FSP) implemented:

1. We can estimate the best response up to an ϵ_t error in round t.
2. We can estimate the opponent's strategy with noisy predictions modeled by M_t in round t.

[14] Leslie and Collins, "Generalized weakened fictitious play"

Estimating the best response

$$b^i_{\epsilon_t}(\pi^{-i}_t + M^i_{t+1})$$

Estimating the best response

$$b^i_{\epsilon_t}(\pi_t^{-i} + M^i_{t+1})$$

By fixing $(\pi_t^{-i} + M^i_{t+1})$, the problem of finding the best response is turned into a single agent utility maximization a.k.a Reinforcement Learning to find an ϵ_t-optimal policy!

Estimating the best response

$$b^i_{\epsilon_t}(\pi_t^{-i} + M^i_{t+1})$$

By fixing $(\pi_t^{-i} + M^i_{t+1})$, the problem of finding the best response is turned into a single agent utility maximization a.k.a Reinforcement Learning to find an ϵ_t-optimal policy!

To get around the exploration requirements of reinforcement learning, SFP[15] used the offline method FQI[16] to learn $b^i_{\epsilon_t}$.

[15] Heinrich, Lanctot, and Silver, "Fictitious self-play in extensive-form games"

[16] Ernst, Geurts, and Wehenkel, "Tree-based batch mode reinforcement learning"

Estimating the opponent's strategy

$$b^i_{\epsilon_t}(\pi_t^{-i} + M^i_{t+1})$$

Estimating the opponent's strategy

$$b^i_{\epsilon_t}(\pi^{-i}_t + M^i_{t+1})$$

To estimate π^{-i}_t, 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 is enough. However, sampled distribution $\hat{\pi}^{-i}$ is a noisy estimation of the true distribution of π^{-i} which is captured by $M^i_t = \frac{1}{\alpha_t}(\hat{\pi}_t^{-i} - \pi^{-i}_t)$.

Estimating the opponent's strategy

Let $\mathcal{A}_i$ be the set of actions available to player i, a set of sampled tuples, (h_t^i, ρ_i^t), where h_t^i is agent i's information set and ρ_i^t is the policy that the agent pursued at this set when this experience was sampled from the dataset. For each tuple (h_t^i, ρ_i^t) the update accumulates each action's weight at the information set:

$$\forall a \in \mathcal{A}(h_t); N(h_t, a) \leftarrow N(h_t, a) + \rho_t(a)$$

$$\forall a \in A(h_t); \hat{\pi}(h_t, a) \leftarrow \frac{N(h_t, a)}{N(h_t)}$$

Results

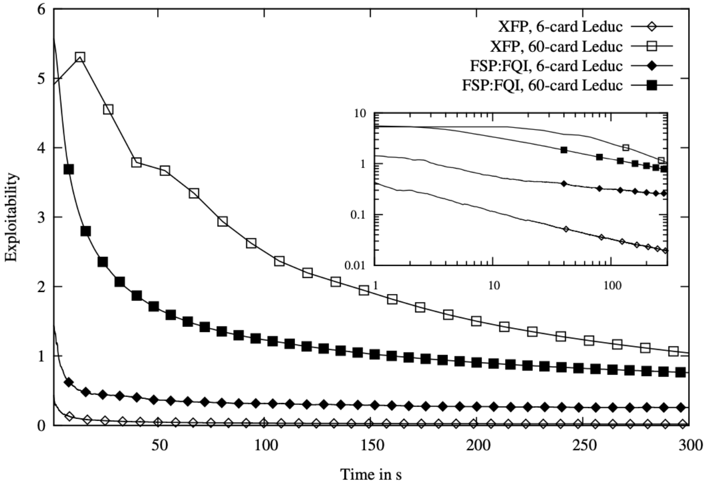


Figure 2. Comparison of XFP and FSP:FQI in Leduc Holdem. The inset presents the results using a logarithmic scale.

Results

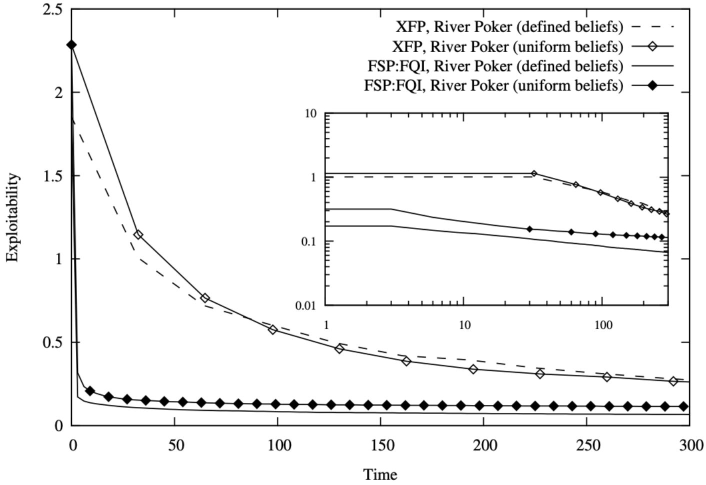


Figure 3. Comparison of XFP and FSP:FQI in River poker. The inset presents the results using a logarithmic scale for both axes.

Neural Self Fictitious Play (NSFP)[17]

Algorithm 1 Neural Fictitious Self-Play (NFSP) with fitted Q-learning

Initialize game Γ and execute an agent via RUNAGENT for each player in the game
function RUNAGENT(Γ)
    Initialize replay memories $\mathcal{M}_{RL}$ (circular buffer) and $\mathcal{M}_{SL}$ (reservoir)
    Initialize average-policy network $\Pi(s, a \mid \theta^{\Pi})$ with random parameters θ^{Π}
    Initialize action-value network $Q(s, a \mid \theta^{Q})$ with random parameters θ^{Q}
    Initialize target network parameters $\theta^{Q'} \leftarrow \theta^{Q}$
    Initialize anticipatory parameter η
    for each episode **do**
        Set policy $\sigma \leftarrow \begin{cases} \epsilon\text{-greedy}\,(Q)\,, & \text{with probability } \eta \\ \Pi, & \text{with probability } 1-\eta \end{cases}$
        Observe initial information state s_1 and reward r_1
        for $t = 1, T$ **do**
            Sample action a_t from policy σ
            Execute action a_t in game and observe reward r_{t+1} and next information state s_{t+1}
            Store transition $(s_t, a_t, r_{t+1}, s_{t+1})$ in reinforcement learning memory $\mathcal{M}_{RL}$
            if agent follows best response policy $\sigma = \epsilon$-greedy (Q) **then**
                Store behaviour tuple (s_t, a_t) in supervised learning memory $\mathcal{M}_{SL}$
            end if
            Update θ^{Π} with stochastic gradient descent on loss
                $\mathcal{L}(\theta^{\Pi}) = \mathbb{E}_{(s,a)\sim\mathcal{M}_{SL}}\left[-\log \Pi(s, a \mid \theta^{\Pi})\right]$
            Update θ^{Q} with stochastic gradient descent on loss
                $\mathcal{L}\left(\theta^{Q}\right) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{M}_{RL}}\left[\left(r + \max_{a'} Q(s', a' \mid \theta^{Q'}) - Q(s, a \mid \theta^{Q})\right)^2\right]$
            Periodically update target network parameters $\theta^{Q'} \leftarrow \theta^{Q}$
        end for
    end for
end function

[17] Heinrich and Silver, "Deep reinforcement learning from self-play in imperfect-information games"

NSFP

Algorithm 1 Neural Fictitious Self-Play (NFSP) with fitted Q-learning

Initialize game Γ and execute an agent via RUNAGENT for each player in the game
function RUNAGENT(Γ)
  Initialize replay memories $\mathcal{M}_{RL}$ (circular buffer) and $\mathcal{M}_{SL}$ (reservoir)
  Initialize average-policy network $\Pi(s, a \mid \theta^{\Pi})$ with random parameters θ^{Π}
  Initialize action-value network $Q(s, a \mid \theta^{Q})$ with random parameters θ^{Q}
  Initialize target network parameters $\theta^{Q'} \leftarrow \theta^{Q}$
  Initialize anticipatory parameter η
  for each episode **do**

$$\text{Set policy } \sigma \leftarrow \begin{cases} \epsilon\text{-greedy}\,(Q)\,, & \text{with probability } \eta \\ \Pi, & \text{with probability } 1-\eta \end{cases}$$

    Observe initial information state s_1 and reward r_1
    for $t = 1, T$ **do**
      Sample action a_t from policy σ
      Execute action a_t in game and observe reward r_{t+1} and next information state s_{t+1}
      Store transition $(s_t, a_t, r_{t+1}, s_{t+1})$ in reinforcement learning memory $\mathcal{M}_{RL}$
      if agent follows best response policy $\sigma = \epsilon$-greedy (Q) **then**
        Store behaviour tuple (s_t, a_t) in supervised learning memory $\mathcal{M}_{SL}$
      end if
      Update θ^{Π} with stochastic gradient descent on loss

$$\boxed{\mathcal{L}(\theta^{\Pi}) = \mathbb{E}_{(s,a)\sim\mathcal{M}_{SL}}\left[-\log \Pi(s, a \mid \theta^{\Pi})\right]}$$

      Update θ^{Q} with stochastic gradient descent on loss

$$\boxed{\mathcal{L}\left(\theta^{Q}\right) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{M}_{RL}}\left[\left(r + \max_{a'} Q(s', a' \mid \theta^{Q'}) - Q(s, a \mid \theta^{Q})\right)^2\right]}$$

      Periodically update target network parameters $\theta^{Q'} \leftarrow \theta^{Q}$
    end for
  end for
end function

Thank You! *:)*