Fictitious Self-Play in Extensive-Form Games

Johannes Heinrich[1], Marc Lanctot[2], David Silver[2]
[1]University College London, [2]Google DeepMind

July 9, 2015

Problem

Learn from self-play in games with imperfect information .

- Games: Multi-agent decision making domains, e.g. poker, politics, security
- Self-play: Agents learn from interaction with each other, without prior knowledge

Extensive-Form Game

Game-theoretic model of sequential interaction of multiple players

- Based on a game tree
- Can represent imperfect (asymmetric, private) information

Fictitious Play

Game-theoretic model of learning in games

- Players repeatedly play a game
- At each iteration each player chooses a best response to their opponents' average behaviour
- Players' average strategies converge to Nash equilibrium in some classes of games, e.g. potential games and two-player zero-sum games

(Brown, 1951)

Fictitious Play

Game-theoretic model of learning in games

- Players repeatedly play a game
- At each iteration each player chooses a best response to their opponents' average behaviour
- Players' average strategies converge to Nash equilibrium in some classes of games, e.g. potential games and two-player zero-sum games

Problem: Almost exclusively studied in normal-form games (tabular, no explicit sequential structure)

(Brown, 1951)

Two algorithms

Fictitious play in extensive-form games

- Computation linear in time and space rather than exponential
- Preserve convergence guarantees

Fictitious Self-Play

- Experiential and sample-based approximation of fictitious play
- Leverages machine learning

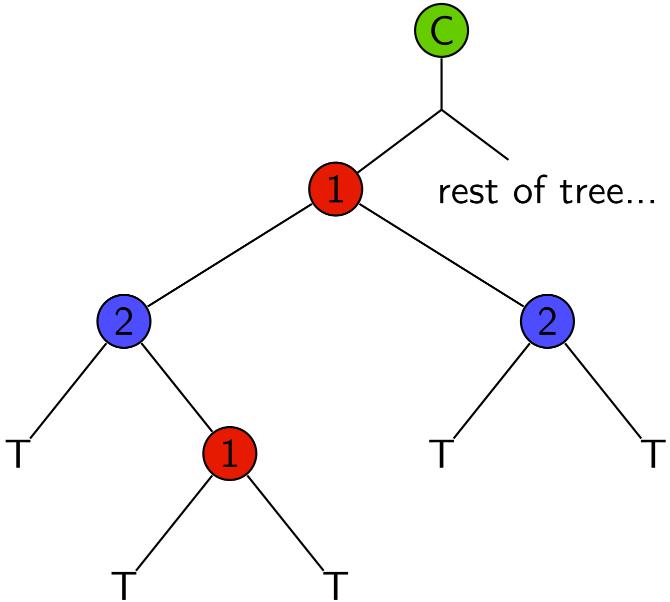
Two steps of fictitious play

At each iteration

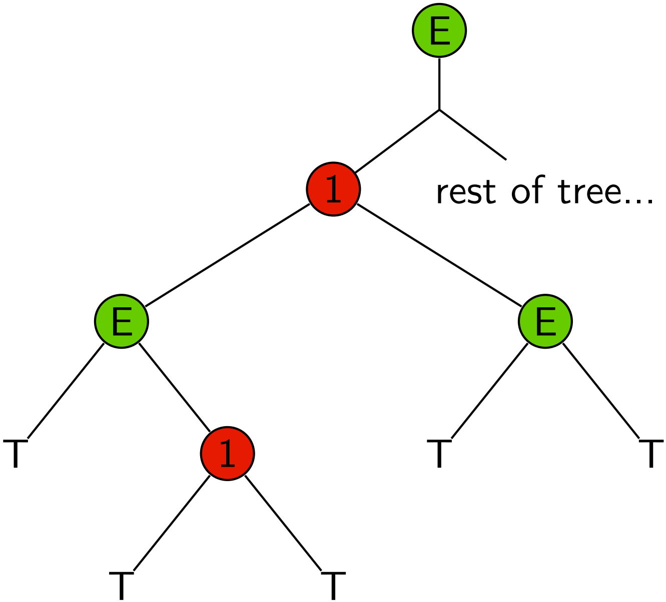
1. Compute a best response to opponents' average strategies
2. Update own average strategy with computed best response

$$\Pi_{k+1} \in \Pi_k + \frac{1}{k+1}\left(\mathrm{BR}\left[\Pi_k\right] - \Pi_k\right)$$

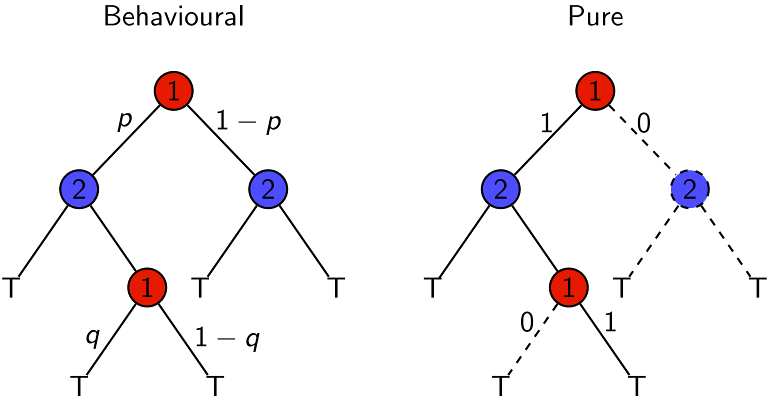
Information state tree



Computing a best response



Strategies

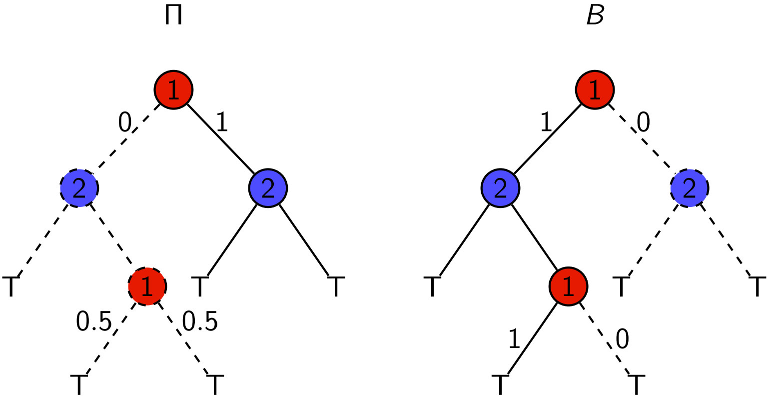


Mixed

pq	$p(1-q)$	$1-p$	0

Aggregating strategies

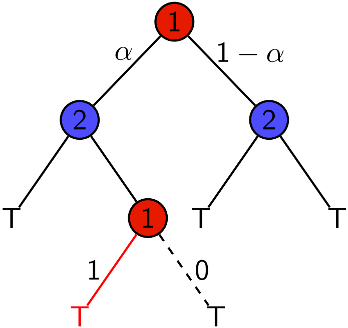
$$(1-\alpha)\overset{\Pi}{\boxed{\begin{array}{c|c|c|c} 0 & 0 & 0.5 & 0.5 \end{array}}} + \alpha \overset{B}{\boxed{\begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \end{array}}}$$



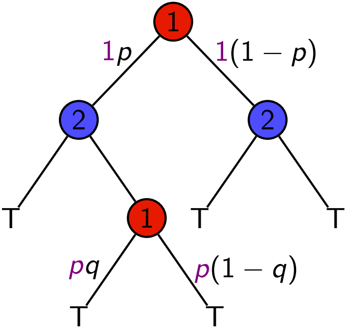
Aggregating strategies

$$(1-\alpha)\Pi + \alpha B$$

α	0	$0.5(1-\alpha)$	$0.5(1-\alpha)$



Realization-equivalence



- Two strategies are **realization-equivalent** iff for any opponent strategies they define the same probability distribution over the states of the game.
- Realization plan:

$$x_\pi(\sigma_u) := \prod_{(u',a)\in\sigma_u} \pi(u', a) \quad \forall u \in \mathcal{U}$$

(Koller et al., 1994; Von Stengel, 1996)

Aggregating behavioural strategies

Lemma

Given

1. π *and* β *two behavioural strategies*
2. Π *and* B *two mixed strategies*
3. Π *and* B *are realization equivalent to* π *and* β

Then

$$\mu(u) = \pi(u) + \alpha \left[\frac{x_\beta(\sigma_u)}{(1-\alpha)x_\pi(\sigma_u) + \alpha x_\beta(\sigma_u)} \right] (\beta(u) - \pi(u)) \quad \forall u \in \mathcal{U}$$

is realization-equivalent to

$$M = \Pi + \alpha\,(B - \Pi)$$

Fictitious play in extensive-form games

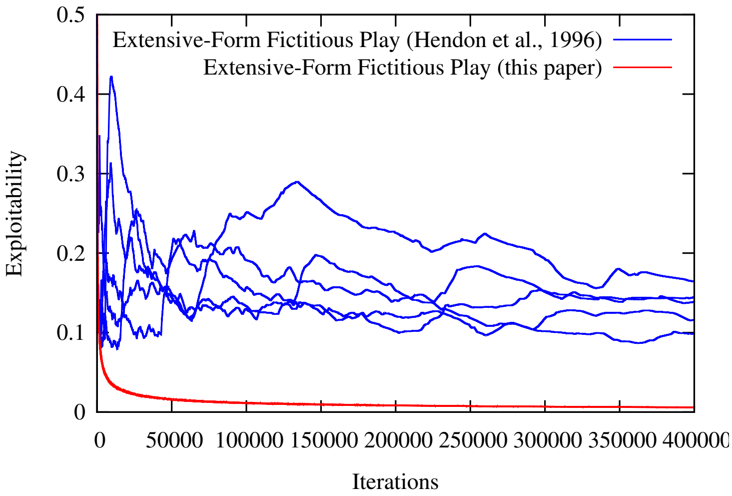


Figure: Learning curves in Leduc Hold'em.

Motivation

- Full-width fictitious play performs computation at all states of the game, no matter whether they are likely to occur
  - Sampling can focus on relevant states
- In big games event a single iteration of full-width fictitious play might be too costly
  - Function approximator could generalise between states
- The state space is larger than information state space
  - Learning agents only operate on their information states

Generalised weakened fictitious play

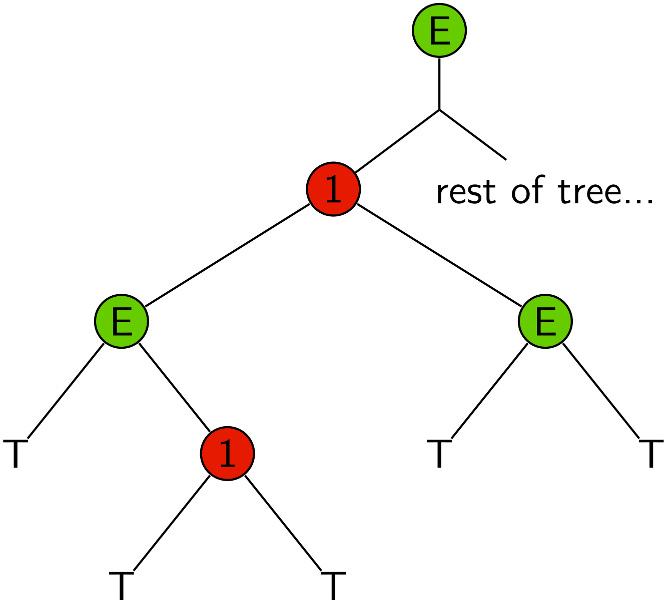
At each iteration

1. Compute an approximate best response to opponents' average strategies
2. Update own average strategy with computed best response, allowing for some kinds of perturbations

$$\Pi_{k+1} \in \Pi_k + \alpha_{k+1} \left(\mathrm{BR}_{\epsilon_{k+1}} [\Pi_k] - \Pi_k + M_{k+1} \right)$$

(Benaïm et al., 2005; Leslie & Collins, 2006)

Learning a best response

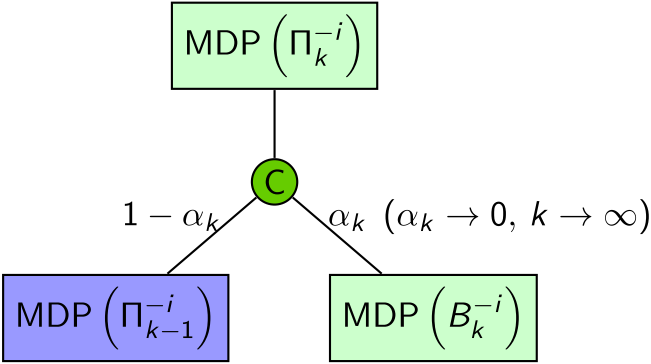


Knowledge transfer

Opponent's update (in a two-player game):

$$\Pi_k^{-i} = (1 - \alpha_k)\Pi_{k-1}^{-i} + \alpha_k B_k^{-i}$$

Thus the MDP defined by Π_k^{-i} has the following structure:



Learning a strategy update

Learn

$$\Pi_k^i = (1-\alpha_k)\Pi_{k-1}^i + \alpha B_k^i,$$

by sampling data (state-action pairs) from

$$\begin{cases} \pi_{k-1}^i & \text{with prob. } 1-\alpha_k \\ \beta_k^i & \text{with prob. } \alpha_k \end{cases}$$

against some fixed, fully mixed opponent sampling policy, e.g. π_{k-1}^{-i}.

Fictitious Self-Play

1. Generate an approximate best response by reinforcement learning from experience
2. Learn a model of own average behaviour by supervised learning from experience
3. Generate experience from self-play, using combinations of $(\pi^i, \pi^{-i}), (\beta^i, \pi^{-i}), (\pi^i, \beta^{-i})$

- Fitted Q-Iteration
- Counting model:

$$\forall a \in \mathcal{A}(u_t) : N(u_t, a) \leftarrow N(u_t, a) + \mathbf{1}_{\{a_t = a\}}$$
$$\forall a \in \mathcal{A}(u_t) : \pi(u_t, a) \leftarrow \frac{N(u_t, a)}{N(u_t)}$$

Fictitious Self-Play in Leduc Hold'em

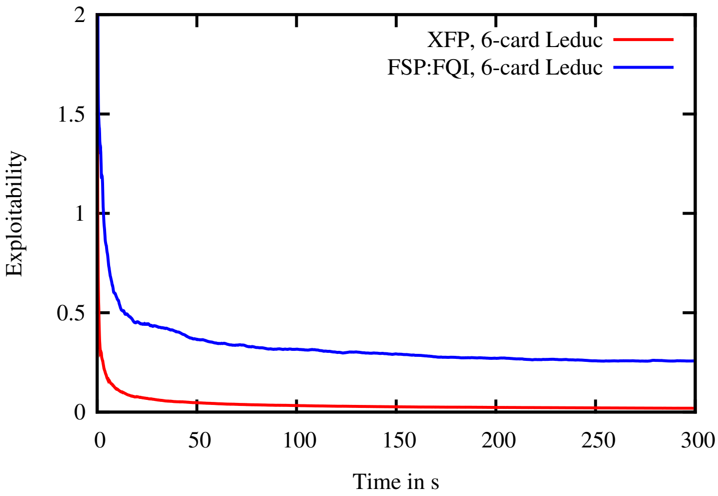


Figure: Learning curves in Leduc Hold'em.

Fictitious Self-Play in Leduc Hold'em

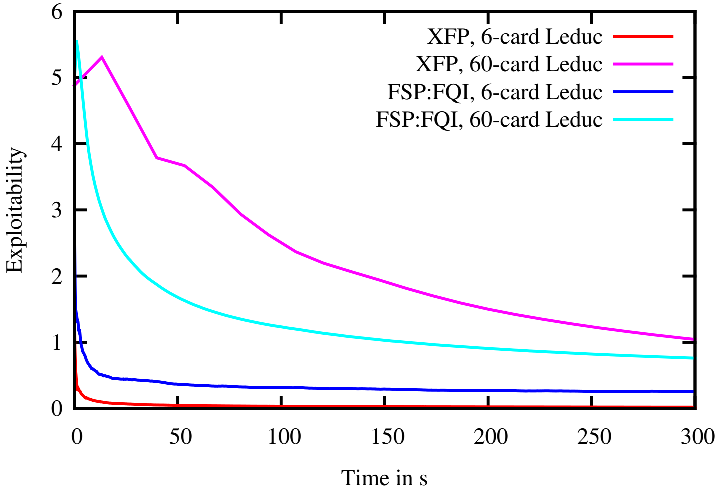


Figure: Learning curves in Leduc Hold'em.

Fictitious Self-Play in River Poker

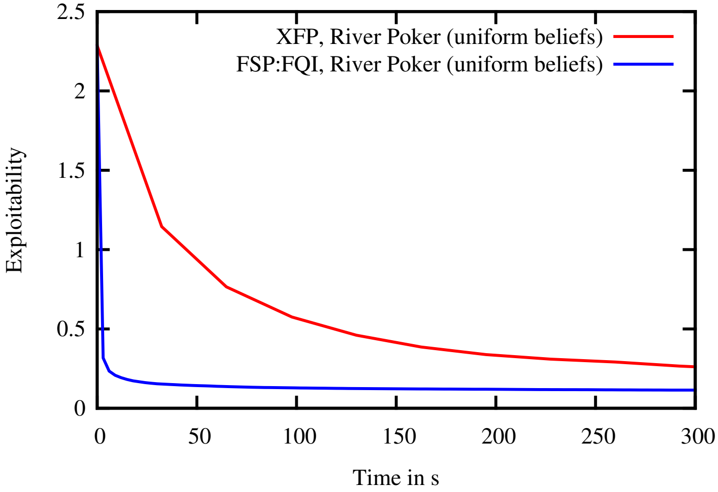


Figure: Learning curves in River Poker.

Fictitious Self-Play in River Poker

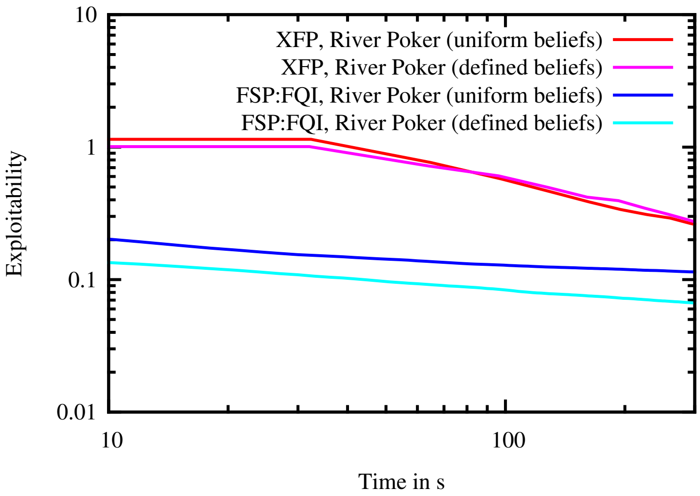


Figure: Learning curves in River Poker.

Summary

- Convergent, full-width fictitious play in extensive-form games
- Fictitious Self-Play is an experiential, sample- and learning-based approach to fictitious play