(Smooth) Fictitious Play for Stochastic Games

Lucas Baudin (with Rida Laraki, Laurent Gourvès, Guillaume Vigeral)
Université Paris-Dauphine

31 January 2023

Outline

1. Learning in Repeated/Stochastic Games
2. Fictitious Play for Repeated Games
3. Q-learning for Reinforcement Learning
4. Combining FP and Q-learning to Learn in Stochastic Games
5. Extension to Unknown Transitions and Perturbed Payoffs

Learning in Repeated/Stochastic Games

Learning in Repeated Games

What is it?

- a procedure that given the history of past rounds, gives an action for the next round
- a dynamic solution concept: learning in repeated games

Learning in Repeated Games

What is it?

- a procedure that given the history of past rounds, gives an action for the next round
- a dynamic solution concept: learning in repeated games

Questions:

- how can such strategies be defined?
- what is the behavior of the dynamics?
- does such a repeated play converge to a (Nash) equilibrium?

Two Widely Studied Learning Procedures

Fictitious Play for Repeated Games

- Brown [1] Robinson [7]
- play a best response to the empirical average of past actions of other players

Two Widely Studied Learning Procedures

Fictitious Play for Repeated Games

- Brown [1] Robinson [7]
- play a best response to the empirical average of past actions of other players

- stochastic games: repeated games with a state variable

Q-Learning for One-Player Stochastic Games

- Watkins [9]
- estimates a table of state-action continuation values

Two Widely Studied Learning Procedures

Fictitious Play for Repeated Games

- Brown [1] Robinson [7]
- play a best response to the empirical average of past actions of other players

- stochastic games: repeated games with a state variable

Q-Learning for One-Player Stochastic Games

- Watkins [9]
- estimates a table of state-action continuation values

- **how can we combine these procedures for multiplayer stochastic games?**

Learning in Stochastic Games

Our 1st paper: based on ideas of Q-Learning and Fictitious Play, we propose a definition of Fictitious Play for multiplayer stochastic games.

Learning in Stochastic Games

Our 1st paper: based on ideas of Q-Learning and Fictitious Play, we propose a definition of Fictitious Play for multiplayer stochastic games.

Convergence results: If all players follow the procedure, then empirical actions converge to:

- the set of stationary Nash equilibria for ergodic, identical-interest stochastic games
- the set of approximate Nash equilibria for ergodic, zero-sum stochastic games.

Our 2nd paper: we extend the definition of FP to smooth action selection for stochastic games with unknown transitions and perturbed payoffs.

Our 2nd paper: we extend the definition of FP to smooth action selection for stochastic games with unknown transitions and perturbed payoffs.

Motivation:

- FP has regret: since it is (almost) deterministic, an other player can take advantage of the procedure
- Smooth FP is known to be "no-regret".

Fictitious Play for Repeated Games

Games

Definition (Game)

$G = (I, (A^i)_{i \in I}, (r^i)_{i \in I})$ where

- I is the finite set of players
- A^i is the finite action set of player i
- $r^i : A \to \mathbb{R}$ is the reward of player i

Nash equilibrium

An action profile where no unilateral deviation are profitable.

Repeated Games

How is a game repeated?

- **sequence of play**: for all steps $n \in \mathbb{N}$
  - every player i plays an action a_n^i
  - every player i receives $r^i(a_n)$

How is a game repeated?

- **sequence of play**: for all steps $n \in \mathbb{N}$
  - every player i plays an action a_n^i
  - every player i receives $r^i(a_n)$
- **discounted payoff**
  - $(1-\delta)\sum_{n=0}^{\infty} \delta^n r^i(a_n)$
    where $\delta \in (0,1)$ is the discount factor

Equilibrium in Repeated Games

The repeated game is a game itself, and has equilibria.

- multiple Nash equilibria: *Folk theorem*

Equilibrium in Repeated Games

The repeated game is a game itself, and has equilibria.

- multiple Nash equilibria: *Folk theorem*
- we are interested in strategies which do not depend on history nor on time, i.e. *stationary strategies and equilibria*

Equilibrium in Repeated Games

The repeated game is a game itself, and has equilibria.

- multiple Nash equilibria: *Folk theorem*
- we are interested in strategies which do not depend on history nor on time, i.e. *stationary strategies and equilibria*
- *lemma:* stationary equilibria are equilibria of the static game

Fictitious Play

- fictitious play is a strategy of the repeated game
- a player plays a best response to the empirical average action of other players

Fictitious Play

- fictitious play is a strategy of the repeated game
- a player plays a best response to the empirical average action of other players

Fictitious Play (Brown [1], Robinson [7])

- empirical average of every player's action:

$$x_n^i = \frac{\sum_{k=0}^{n} a_k^i}{n}$$

- action selection:

$$a_{n+1}^i \in \mathrm{BR}(x_n^{-i}) := \arg\max_{b^i \in A^i} r^i(b^i, x_n^{-i})$$

Fictitious Play

- fictitious play is a strategy of the repeated game
- a player plays a best response to the empirical average action of other players

Fictitious Play (Brown [1], Robinson [7])

- empirical average of every player's action:

$$x_n^i = \frac{\sum_{k=0}^{n} a_k^i}{n}$$

- action selection:

$$a_{n+1}^i \in \mathrm{BR}(x_n^{-i}) := \arg\max_{b^i \in A^i} r^i(b^i, x_n^{-i})$$

- *Remark:* every player plays assuming that other players are stationary

Convergence

Convergence

If all players use fictitious play, then the average actions converge to the set of stationary Nash equilibria for several classes of games:

- zero-sum games (Brown [1], Robinson [7])
- potential games (Monderer and Shapley [6])...

Q-learning for Reinforcement Learning

Stochastic Games (Definition)

Definition (Stochastic Game)

$G = (S, I, (A^i)_{i \in I}, (r^i_s)_{i \in I, s \in S}, (P_s)_{s \in S})$

- S is a finite state space
- A^i is the action set of player i
- $r^i_s : A \to \mathbb{R}$ is the stage reward
- $P_s : A \to \Delta(S)$ is the transition probability map.

Stochastic Games (Definition)

Definition (Stochastic Game)

$G = (S, I, (A^i)_{i \in I}, (r^i_s)_{i \in I, s \in S}, (P_s)_{s \in S})$

- S is a finite state space
- A^i is the action set of player i
- $r^i_s : A \to \mathbb{R}$ is the stage reward
- $P_s : A \to \Delta(S)$ is the transition probability map.

We focus on two classes of games:

- identical interest: $r^i_s = r_s$
- zero sum: $r^1_s = -r^2_s$

- ergodic: every state s' is reached from any state s with positive probability for any sequence of actions in a finite time

Playing Stochastic Games

How to play stochastic games?

- initial state s_0
- for all steps $n \in \mathbb{N}$, the system is in s_n:
  - every player i plays an action a_n^i
  - every player i receives $r_{s_n}^i(a_n)$
  - new state $s_{n+1} \sim P_{s_n}(a_n)$

Equilibria of Stochastic Games

- **discounted payoff**
  - $(1-\delta)\sum_{n=0}^{\infty}\delta^n r^i_{s_n}(a_n)$
    where $\delta \in (0,1)$ is the discount factor

Equilibria of Stochastic Games

- **discounted payoff**
  - $(1-\delta)\sum_{n=0}^{\infty}\delta^n r^i_{s_n}(a_n)$
    where $\delta \in (0,1)$ is the discount factor
- **equilibria**: a stochastic game has equilibria
- we are interested in the convergence of our procedures to stationary equilibria [2]
- *lemma:* a player has an optimal stationary strategy if other players are stationary

Q-Learning: the One-Player Case

- Q-Learning: a procedure that updates a Q function
- $Q(s, a)$ = continuation payoff in s when a is played

Q-Learning: the One-Player Case

- Q-Learning: a procedure that updates a Q function
- $Q(s, a) =$ continuation payoff in s when a is played

Q-learning (Watkins [9])

At every step n, if the system is in s_n and a_n is played, then:

$$Q_{n+1}(s_n, a_n) \leftarrow Q_n(s_n, a_n) + \gamma \left(R_{n+1} + \delta \max_a Q_n(s_{n+1}, a) - Q_n(s_n, a_n) \right)$$

where $R_{n+1} = (1 - \delta) r_{s_n}(a_n)$ and γ is the update step.

- convergence with **one player** when the environment is stationary and the update step decreasing
- **problem:** in multiplayer stochastic games, other player actions are not stationary

Combining FP and Q-learning to Learn in Stochastic Games

Combining FP and Q-learning

Inspired by Leslie et al. [5]; Sayin et al. [8].

- two sets of variables
  - estimate u_s of the continuation payoff starting from a state s
  - estimate x_s^i of other player i strategy in state s that will be used by other players

Combining FP and Q-learning

Inspired by Leslie et al. [5]; Sayin et al. [8].

- two sets of variables
  - estimate u_s of the continuation payoff starting from a state s
  - estimate x_s^i of other player i strategy in state s that will be used by other players
- variables are updated at every step: sequence $(u_{s,n}, x_{s,n})$.

Auxiliary Game

We define an auxiliary game using a vector u of continuation payoffs.

Definition (Auxiliary Game)

- one-shot, static game parameterized by a vector u
- actions A
- payoff functions:

$$f_{s,u}(a) = (1-\delta)r_s(a) + \delta \sum_{s' \in S} P_{ss'}(a) u_{s'}$$

Auxiliary Game

We define an auxiliary game using a vector u of continuation payoffs.

Definition (Auxiliary Game)

- one-shot, static game parameterized by a vector u
- actions A
- payoff functions:

$$f_{s,u}(a) = (1-\delta) r_s(a) + \delta \sum_{s' \in S} P_{ss'}(a) u_{s'}$$

Remark: $f_{s,u}$ is extended to mixed action profiles

Auxiliary Game

We define an auxiliary game using a vector u of continuation payoffs.

Definition (Auxiliary Game)

- one-shot, static game parameterized by a vector u
- actions A
- payoff functions:

$$f_{s,u}(a) = (1-\delta) r_s(a) + \delta \sum_{s' \in S} P_{ss'}(a) u_{s'}$$

Remark: $f_{s,u}$ is extended to mixed action profiles

Remark: it corresponds to a one-shot game whose payoff is the instantaneous payoff of the stochastic games + the estimate of the continuation payoff in u.

Fictitious Play for Stochastic Games

FP for stochastic games for all players

- action selection: a best response in the auxiliary game parameterized by u_n to empirical action $x_{s,n}^{-i}$
- update of u_n: towards the payoff in the auxiliary game $f_{s,u_n}(x_{s,n})$
- update of $x_{s,n+1}^{i}$: empirical action of player i in state s

Fictitious Play for Stochastic Games

FP for stochastic games for all players

- action selection: a best response in the auxiliary game parameterized by u_n to empirical action $x_{s,n}^{-i}$
- update of u_n: towards the payoff in the auxiliary game $f_{s,u_n}(x_{s,n})$
- update of $x_{s,n+1}^{i}$: empirical action of player i in state s

Fictitious Play for Stochastic Games

FP for stochastic games for all players

- $\forall s\ u_{s,n+1} - u_{s,n} = \dfrac{\beta}{n+1}\left(f_{s,u_n}(x_{s,n}) - u_{s,n}\right)$
- $a^i_{n+1} \in \arg\max_{b^i \in A^i} f_{s_{n+1},u_{n+1}}(b^i, x^{-i}_{s_{n+1},n})$
- $x^i_{s,n+1} = \dfrac{\sum_{k=0}^{n+1} 1_{s_k = s} a^i_n}{s^\sharp_n}$

where $s^\sharp_n = \sharp\{i \mid 0 \le i \le n \wedge s_i = s\}$ and $\beta > 0$

Fictitious Play for Stochastic Games

FP for stochastic games for all players

- $\forall s \; u_{s,n+1} - u_{s,n} = \frac{\beta}{n+1} \left(f_{s,u_n}(x_{s,n}) - u_{s,n} \right)$
- $a^i_{n+1} \in \arg\max_{b^i \in A^i} f_{s_{n+1},u_{n+1}}(b^i, x^{-i}_{s_{n+1},n})$
- $x^i_{s,n+1} - x^i_{s,n} = \frac{1_{s_{n+1}=s}}{s^\sharp_{n+1}} (a^i_{n+1} - x^i_{s,n})$

Convergence

Set-up: all players use FP, we look at empirical actions.

Convergence

Set-up: all players use FP, we look at empirical actions.

Theorem (convergence of FP in i.i. stochastic games)

For identical-interest ergodic stochastic games, FP for stochastic games converges to the set of stationary Nash equilibrium.

Convergence

Set-up: all players use FP, we look at empirical actions.

Theorem (convergence of FP in i.i. stochastic games)

For identical-interest ergodic stochastic games, FP for stochastic games converges to the set of stationary Nash equilibrium.

Theorem (convergence of FP in z.s. stochastic games)

For zero-sum ergodic stochastic games, FP for stochastic games converges to the set of stationary $A\beta$-Nash equilibrium where $A > 0$ does not depend on β.

Synchronicity

- FP is updating empirical actions for the **current state** and continuations payoff for **all states**
- we now define other procedures where the variables are updated for **all the states** or only for the **current state**

Synchronicity

- FP is updating empirical actions for the **current state** and continuations payoff for **all states**
- we now define other procedures where the variables are updated for **all the states** or only for the **current state**

Synchronicity

- FP is updating empirical actions for the **current state** and continuations payoff for **all states**
- we now define other procedures where the variables are updated for **all the states** or only for the **current state**

Synchronous FP

- $u_{s,n+1} - u_{s,n} = \frac{1}{n+1}\left(f_{s,u_n}(x_{s,n}) - u_{s,n}\right)$
- $a^i_{s,n+1} \in \arg\max_{b^i \in A^i} f_{s,u_{n+1}}(b^i, x^{-i}_{s,n})$
- $x^i_{s,n+1} - x^i_{s,n} = \frac{1}{n+1}\left(a^i_{s,n+1} - x^i_{s,n}\right)$

Fully-asynchronous FP

- $u_{s,n+1} - u_{s,n} = \frac{1_{s_{n+1}=s}}{s^\sharp_{n+1}}\left(f^i_{s,u_n}(x_{s,n}) - u^i_{s,n}\right)$
- $x^i_{s,n+1} - x^i_{s,n} = \frac{1_{s_{n+1}=s}}{s^\sharp_{n+1}}\left(a^i_{n+1} - x^i_{s,n}\right)$

Convergence

Theorem (convergence of FP in i.i. stochastic games)

For identical interest ergodic stochastic games, synchronous FP for stochastic games converges to the set of stationary Nash equilibrium.

Fully-asynchronous FP also converges if $\delta < 1/|S|$.

Proof

idea:

- first, define analogous continuous-time systems

Proof

idea:

- first, define analogous continuous-time systems
- second, study the convergence in these continuous-time systems

Proof

idea:

- first, define analogous continuous-time systems
- second, study the convergence in these continuous-time systems
- third, use the stochastic approximation framework to deduce results in discrete time

Proof (2)

In continuous time, we get a best-response dynamics:

Synchronous Best-Response Dynamics

$$\begin{cases} \dot{u}_s = f_{s,u}(x) - u_s \\ \dot{x}_s^i \in \mathrm{BR}_{u,s}(x_s^{-i}) - x_s^i \end{cases}$$

Proof (3)

continuous: $\frac{dx}{dt} \in F(x)$

discrete-time: $x_{n+1} - x_n \in \gamma_n F(x_n)$

Proof (3)

$$\text{continuous: } \frac{dx}{dt} \in F(x)$$

$$\text{discrete-time: } x_{n+1} - x_n \in \gamma_n F(x_n)$$

Stochastic Approximations

- if $F : \mathbb{R}^k \rightrightarrows \mathbb{R}^k$ is a Marchaud map
- γ_n such that $\gamma_n \geq 0$, $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$

These two class of sets are equal:

- internally chain transitive sets for $\frac{dx}{dt} \in F(x)$
- limit sets of $x_{n+1} - x_n \in \gamma_n F(x_n)$

Extension

- **idea:** update the u_n vector slower than the x_n vectors (Leslie et al. [4], Sayin et al. [8])

Extension

- **idea:** update the u_n vector slower than the x_n vectors (Leslie et al. [4], Sayin et al. [8])

FP for Stochastic Game

- $s_n^\sharp = \sharp\{k \mid 0 \le k \le n \wedge s_k = s\}$
- $u^i_{s,n+1} - u^i_{s,n} = \frac{1_{s_{n+1}=s}}{\alpha(s_n^\sharp)} \left(f^i_{s,u_i}(x_{s,n}) - u^i_{s,n} \right)$
- $x^i_{s,n+1} - x^i_{s,n} = \frac{1_{s_{n+1}=s}}{s_n^\sharp} \left(a^i_n - x^i_{n,s} \right)$
- $a^i_{n+1} \in \arg\max f^i_{u_{n+1},s_{n+1}}(x^i_{s,n+1})$

- **idea:** extend the proofs to other classes of games

Extension to Unknown Transitions and Perturbed Payoffs

Smooth Fictitious Play

Fictitious Play

$$a^i_{n+1} \in \mathrm{BR}(x^{-i}_n) := \arg\max_{b^i \in A^i} r^i(b^i, x^{-i}_n)$$

Smooth Fictitious Play

Fictitious Play

$$a^i_{n+1} \in \mathrm{BR}(x^{-i}_n) := \arg\max_{b^i \in A^i} r^i(b^i, x^{-i}_n)$$

Smooth Fictitious Play Fudenberg and Levine [3]

$$a^i_{n+1} \sim \mathrm{SBR}(x^{-i}_n) := \arg\max_{\sigma^i \in \Delta(A^i)} r^i(\sigma^i, x^{-i}_n) + \epsilon h^i(\sigma^i, x^{-i}_n)$$

Smooth Fictitious Play

Fictitious Play

$$a^i_{n+1} \in \mathrm{BR}(x^{-i}_n) := \arg\max_{b^i \in A^i} r^i(b^i, x^{-i}_n)$$

Smooth Fictitious Play Fudenberg and Levine [3]

$$a^i_{n+1} \sim \mathrm{SBR}(x^{-i}_n) := \arg\max_{\sigma^i \in \Delta(A^i)} r^i(\sigma^i, x^{-i}_n) + \epsilon h^i(\sigma^i, x^{-i}_n)$$

Regularizer

- $h^i : \Pi_{j \in I} \Delta(A^j) \mapsto \mathbb{R}^+$, smooth, strictly concave in σ^i
  $\|\nabla h^i\| = +\infty$ on the boundary of $\Delta(A^i)$
- $\epsilon > 0$

Why Smooth Best-Response?

- SFP has the no-regret property while FP has not

Why Smooth Best-Response?

- SFP has the no-regret property while FP has not
- Every action is played infinitely often

Smooth Fictitious Play for Stochastic Games

Our definition of SFP in stochastic games:

SFP for Stochastic Games (known payoff and transition)

$$\begin{cases} u_{s,n+1} - u_{s,n} = \frac{\beta}{n+1}\left(f_{s,u_n}(x_{s,n}) - u_{s,n}\right) \\ a^i_{n+1} \sim \underset{\sigma^i \in \Delta(A^i)}{\arg\max}\, f_{s_{n+1},u_{n+1}}(\sigma^i, x^{-i}_n) + \epsilon h(\sigma^i, x^{-i}_n) \\ x^i_{s,n+1} = \frac{\sum_{k=0}^{n+1} 1_{s_k=s} a^i_k}{s^\sharp_{n+1}} \end{cases}$$

Results

Theorem

SFP for stochastic games converges to

- *the set of regularized Nash equilibrium for identical-interest stochastic games*
- *the set of $M\beta$ regularized Nash equilibria for zero-sum stochastic games.*

SFP with Unknown Transitions and Perturbed Payoffs

Unknown transitions: P_s is unknown but states are observed

SFP with Unknown Transitions and Perturbed Payoffs

Unknown transitions: P_s is unknown but states are observed

Perturbed payoffs: r_s^i are unknown and $E[R_n^i] = r_{s_n}^i(a_n)$

SFP with Unknown Transitions and Perturbed Payoffs

Unknown transitions: P_s is unknown but states are observed

Perturbed payoffs: r_s^i are unknown and $E[R_n^i] = r_{s_n}^i(a_n)$

- $\hat{f}_{s,u_n}(\sigma_s) = (1-\delta)\hat{r}_s(\sigma_s) + \delta \hat{P}_s(\sigma_s) \cdot u_n$
- with $\hat{r}_s$ and $\hat{P}_s$ average vectors of past payoffs and transitions

Results and Proofs

- same results as in the known transitions and payoffs case

Results and Proofs

- same results as in the known transitions and payoffs case
- proofs: uses the continuous-time smooth best-response dynamics

Conclusion

Our results

- procedures to play stochastic games
- convergence of the procedures for identical-interest and zero-sum ergodic stochastic games
- convergence of a generalized continuous-time system

Conclusion

Future work

- other classes of games
- different update steps for x_n and u_n
- suppose less coordination between players: different update steps, different priors

References

[1] George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

[2] A. M. Fink. Equilibrium in a stochastic n-person game. *Hiroshima Mathematical Journal*, 28(1), January 1964. ISSN 0018-2079. doi: 10.32917/hmj/1206139508.

[3] Drew Fudenberg and David K. Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5-7):1065–1089, July 1995. ISSN 01651889. doi: 10.1016/0165-1889(94)00819-4.

[4] David S Leslie, Steven Perkins, and Zibo Xu. Best-response Dynamics in Zero-sum Stochastic Games. page 34, April 2018.

[5] David S. Leslie, Steven Perkins, and Zibo Xu. Best-response dynamics in zero-sum stochastic games. *Journal of Economic Theory*, 189:105095, September 2020. ISSN 00220531. doi: 10.1016/j.jet.2020.105095.

[6] Dov Monderer and Lloyd S. Shapley. Fictitious Play Property for Games with Identical Interests. *Journal of Economic Theory*, 68(1):258–265, January 1996. ISSN 00220531. doi: 10.1006/jeth.1996.0014.

[7] Julia Robinson. An Iterative Method of Solving a Game. *The Annals of Mathematics*, 54(2):296, September 1951. ISSN 0003486X. doi: 10.2307/1969530.

[8] Muhammed O. Sayin, Francesca Parise, and Asuman Ozdaglar. Fictitious play in zero-sum stochastic games. *SIAM Journal on Control and Optimization*, 60(4):2095–2114, 2022. doi: 10.1137/21M1426675.

[9] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Oxford, 1989.