Fictitious Play in Markov Games with Single Controller

Muhammed O. Sayin, Dep. Electrical & Electronics Eng., Bilkent University, Ankara, Turkey
Kaiqing Zhang, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA
Asuman Ozdaglar, Dep. Electrical Eng. & Computer Sci., Massachusetts Institute of Technology, Cambridge, MA

Frontiers in AI:

Multi-agent Learning in **Dynamic** Environments

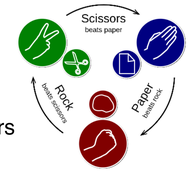
Focus:

- **Individual objective** optimization (beyond equilibrium seeking)
- **Dynamic** environments (beyond repeated games)
- Systematic guarantees

Game Theory:

Nash Equilibrium: Players do not have any incentive to change their strategies unilaterally.

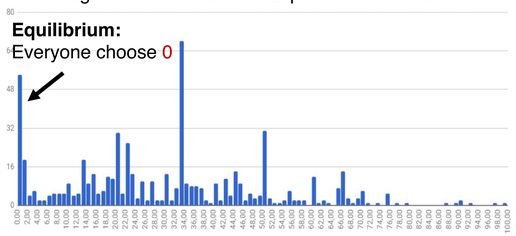
Always exists in **finite** games if players can randomize their actions.



How **predictive** is a Nash equilibrium?

Let's play a game!
1. Everyone will pick an integer between [0,100] simultaneously.
2. The one closest to the 2/3 of the average will share the price.

Histogram of Choices in an Experiment of the Game



Humans are **NOT** necessarily seeking for an equilibrium!

Self-interested players reach to an equilibrium if they play the same game **repeatedly** and receive feedback

Fictitious Play:

- **Model the opponent** as playing a **stationary** strategy
- **Form a belief** about the **opponent strategy**
- **Respond** to the belief by taking a **greedy** best response

Convergence Results for Fictitious Play in **Repeated Games with Self-Play:**

Spectrum of Games

Opposite Goals: + **Zero-sum** ⟷ Common Goals: + **Identical-payoff** + Potential games

What about **BEYOND** the games with repeated play?

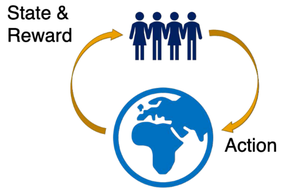
Markov Games (a.k.a. Stochastic Games):

- Multi state environment
- Markovian state transitions
- Markovian rewards
- **Objective:** discounted sum of rewards collected over infinite horizon

$$\langle S, A, r, p, \gamma \rangle$$

Single controller i

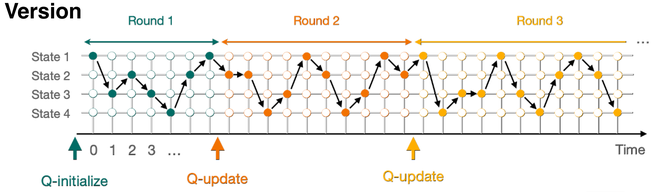
$$p(s'|s,a) = p(s'|s,a^i), \forall (s,a,s')$$



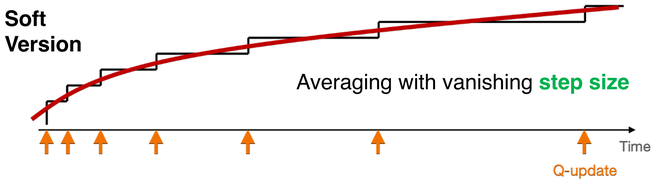
Challenges	Solutions
Trade-off between now and (**ambiguous**) future	(Standard) Define **Q-function** for state-action values
Non-stationarity of how the future is perceived	**(Key Idea)** Two-timescale learning
Deviation from the identical-interest structure in the stage games	**(Key Idea)** Single-controller Markov games for **strategic equivalence to identical interest games**!

Two-timescale Learning

Hard Version



Soft Version



Two-timescale Fictitious Play:

- **Model the opponent** as playing a **stationary** strategy specific to each state
- **Form a belief** about the **opponent strategy** (**Large** steps)
- **Form a belief** about the **Q-function** (**Small** steps)
- **Respond** to the beliefs by taking a **greedy** best response

Player i

For each state s

$$\pi_s^j \leftarrow \pi_s^j + \alpha_k\left(a_s^j - \pi_s^j\right), \forall j \neq i$$
$$a_s^i \in \underset{a^i}{\operatorname{argmax}} \left\{\mathrm{E}_{a^{-i}\sim\pi_s^{-i}}\left[Q_s^i(a^i, a^{-i})\right]\right\}$$
$$Q_s^i \leftarrow Q_s^i + \beta_k\left(R_s + \gamma \mathrm{E}_{s'}[v_{s'}^i] - Q_s^i\right)$$
$$v_s^i \leftarrow \max_{a^i} \left\{\mathrm{E}_{a^{-i}\sim\pi_s^{-i}}\left[Q_s^i(a^i, a^{-i})\right]\right\}$$

Two-timescale step-sizes: $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$

Theorem:

Given a **multi-player identical-interest discounted Markov game**, suppose that players follow the two-timescale fictitious play dynamics with non-summable vanishing step sizes. If each state gets visited infinitely often, then we have the beliefs π and Q **converge to a stationary equilibrium** and the associated Q-function **almost surely**.

Corollary (Potential-game-like extension):

Under the conditions in Theorem, suppose that $r^j(s,\tilde{a}^j,a^{-j}) - r^j(s,a) = r^i(s,\tilde{a}^j,a^{-j}) - r^i(s,a)$ for all (s,a), $\tilde{a}^j$ and $j \neq i$ given that player i is the single controller. Then, we have the beliefs π and Q **converge to a stationary equilibrium** and the associated Q-function **almost surely**.

Spectrum of **Markov** Games

Opposite Goals: + **Zero-sum [1]** ⟷ Common Goals: + **Identical-payoff w/ single controller [This paper]**

Sketch of the Proof:

Show **accumulated monotonicity-like condition**:

$$\liminf_{k_1\to\infty} \inf_{k_2\geq k_1} \left\{Q_{k_2}^i(s,a) - Q_{k_1}^i(s,a)\right\} \geq 0,$$

implying converges since Q-functions are bounded.

Alternatively, define

$$\underline{u} = \min_{(s,a)}\left\{r + \gamma \mathrm{E}_{s'}\left[\mathrm{E}_{a'\sim\pi_k(s')}[Q_k^i(s',a')]\right] - Q_k^i\right\}$$

And show $\liminf_{k_1\to\infty} \inf_{k_2\geq k_1} \sum_{k=k_1}^{k_2} \beta_k \underline{u}_k^i \geq 0.$

We can show it because $\underline{u}_k^i$ satisfies

$$\underline{u}_{k+1}^i \geq \underline{u}_k^i(1-(1-\gamma)\beta_k) + e_k$$

Absolutely summable error term (for **single controller** case)

Given that Q-function estimates converge, we can show they indeed converge to the Q-function associated with an equilibrium. ■

[1] M. O. Sayin, F. Parise, and A. Ozdaglar, "Fictitious play in zero-sum stochastic games," *SIAM J. Cont. Opt.*, in print.

July 11-15, 2022, The 23rd ACM Conference on Economics and Computation (EC'22), the University of Colorado Boulder, CO