Independent Learning Dynamics for Stochastic Games: Convergence and Finite-Time Analysis

Asu Ozdaglar

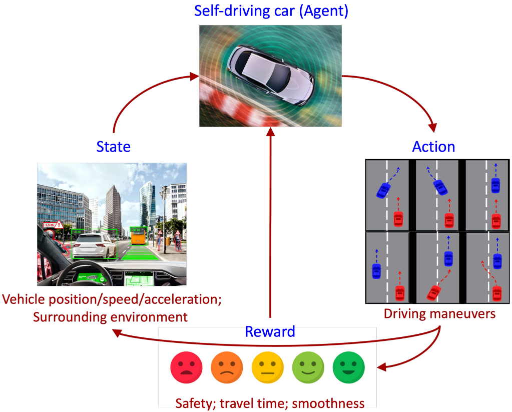
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Joint work with
Kaiqing Zhang (Maryland) and Chanwoo Park (MIT),
Muhammed O. Sayin (Bilkent) and Francesca Parise (Cornell),
Zaiwei Chen (Caltech), Eric Mazumdar (Caltech), and Adam Wierman (Caltech)

ELLIIT Focus Period on Network Dynamics and Control
September, 2023

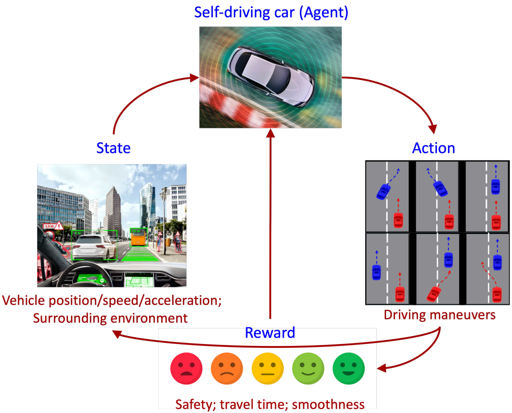
Reinforcement Learning

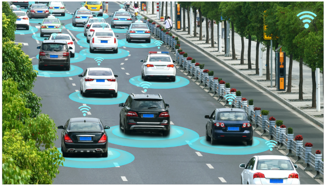
- Reinforcement learning (RL) has emerged as the backbone of many artificial intelligence (AI) problems, where autonomous agents have to make sequential decisions in unknown dynamic environments.



Reinforcement Learning

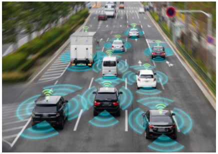
- Reinforcement learning (RL) has emerged as the backbone of many artificial intelligence (AI) problems, where autonomous agents have to make sequential decisions in unknown dynamic environments.





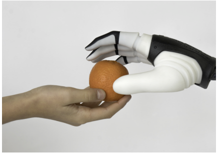
Multi-Agent Reinforcement Learning

- In fact, many more AI systems involve multi-agent dynamic settings:









- Further advances critically depend on analyzing multi-agent interactions, decisions and learning in dynamic environments.

Nash Equilibrium and Learning in Games

- Nash Equilibrium (NE) – a remarkably powerful tool for understanding multi-agent interactions.
- Most economists and computer scientists have come to think of NE as arising not from introspection and calculation, but rather from some non-equilibrium adaptive process of learning [Fudenberg and Levine 16].

Multi-Agent Learning in Static and Dynamic Games

- One of the best studied models of learning is fictitious play (FP):
  - Myopic agents estimate opponent strategy using past play.
  - They use a best-response type action (using their stage payoff) against this estimate.
- Large literature in economics and game theory on convergence of fictitious play for repeated play of static games [Robinson 51], [Monderer and Shapley 96], [Fudenberg and Kreps 93], [Fudenberg and Levine 95].
- Despite its importance, there is limited progress on multi-agent learning in dynamic environments.
- **Key challenge**: Estimating decision rules of other adaptive agents in changing non-stationary environments.
  - These challenges multiplied in the (model-free) RL setting when a dynamic model of the environment (i.e., transition probabilities and payoff functions) is unknown.

Classical Results for Learning in Dynamic Games

Mostly computational in nature and for zero-sum:

- [Shapley 53]:
  - Defined stochastic games (extends strategic form games to dynamic environments and MDPs to competitive situations).
  - Minimax value-iteration (VI) algorithm to compute value functions in zero-some stochastic games.
  - It converges due to the γ-contracting property of the VI operator.
- [Littman 94]:
  - Q-learning in stochastic games, without the model.
  - Extended in [Littman and Szepesvari 96], [Hu and Wellman 03], [Bowling 05].

Recent Results

Two strands of recent literature on multi-agent dynamic learning:

Centralized Learning: Centralized controller that jointly optimizes all agent policies [Perolat et al. 15], [Sidford et al. 19], [Bai, Jin 20], [Shah et al. 20], [Zhang et al. 20].

Decentralized/independent learning: Agents optimize their own payoff given their observations and beliefs.

- **Challenges of independent learning**: Negative non-convergent results due to non-stationarity [Condon 90], [Tan 93], [Claus and Boutilier 98].

Most relevant to our work:

- **Zero-sum Stochastic Games:**
  - [Daskalakis et al. 20] Policy gradient methods: coordination between agents' learning rates.
  - [Leslie et al. 20] Continuous-time best-response dynamics, a common continuation payoff for all players – updated at a slower speed.
- **Potential Stochastic Games:**
  - [Leonardos et al. 21][Zhang et al. 21][Fox et al. 22] Policy gradient methods: algorithmic approaches for equilibrium computation.

Question of Interest

Open question 1: Can we identify reasonable and independent learning dynamics that converge to NE for stochastic games?

- Reasonable: Agents acting in their individual interest.
- Independent: No coordination among agents.

Open question 2: Can we provide finite sample guarantees for best-response type dynamics for stochastic games (even matrix games)?

Our Results - Multi-agent Learning Made Simple

- We develop simple learning rules based on FP-type dynamics that are fully decentralized and independent.
  - Convergence for zero-sum stochastic games [Sayin, Parise, Ozdaglar 21], [Sayin*, Zhang*, Leslie, Başar, Ozdaglar 21]
  - Finite-time and payoff-based analysis for zero-sum stochastic games [Chen, Zhang, Mazumdar, Ozdaglar, Wierman 23]
- We conclude with a new tractable model of multi-player networked Markov games [Park, Zhang, Ozdaglar 23].

Main ideas:

- Two-timescale learning, but only at the individual agent level.
  - Each agent is simultaneously estimating the empirical distribution of others' actions/strategies and his own continuation payoff.
  - Two-timescale here refers to empirical distribution updated more frequently than underlying estimate of the payoff functions.
- For finite-time analysis (and payoff-based dynamics): doubly-smoothed best response dynamics with estimation of local payoff functions.

Model

Stochastic Game

- An n-player stochastic game $\langle S, \{A^i\}_{i\in[n]}, \{r^i\}_{i\in[n]}, p, \gamma\rangle$.
- S is the set of finitely many states.
- A^i is the set of finitely many actions that player i can take at state s. ($\Delta(A^i)$ denotes the set of probability distributions over the set A^i).
- $A = \prod_i A^i$ denotes the set of action profiles $a = [a^i]_{i\in[n]}$.
- $r^i(s, a)$ denotes the stage payoff of player i at state s and action profile a.
- Players take action a at state $s \in S$, and the state transitions to $\tilde{s}$ according to $p(\tilde{s}|s, a)$.
- $\gamma \in [0, 1)$ is the discount factor.

Model

Equilibrium

- We focus on stationary Markov strategies (a mixed strategy per state).
- Let $\pi^i : S \to \Delta(A)$ with $\pi^i(s) \in \Delta(A^i)$ denote the (mixed) strategy of player i at state s and $\pi = (\pi^i)_{i\in[n]}$ denote the strategy profile.
- We define the expected payoff (value) function of player i as

$$v^i(s;\pi) := \mathbb{E}_{a_k\sim\pi(s_k)}\left\{\sum_{k=0}^{\infty}\gamma^k r^i(s_k,a_k)\middle| s_0 = s\right\},$$

where $\{s_k\}_{k\geq 0}$ is a stochastic process. We use $v^i(\pi) = \mathbb{E}_{s\sim p_0}\left\{v^i(s;\pi)\right\}$.

Definition (Nash Equilibrium)

A strategy profile π_* is a (Nash) equilibrium provided that

$$v^i(\pi_*) \geq v^i(\pi^i,\pi_*^{-i}) \qquad \textit{for all } \pi^i, \quad \textit{and all } i.$$

The value $v^i(\pi_*)$ represents the equilibrium value function of player i.

Model

Value function characterization

- Using one-stage deviation principle (multi-agent extension of Bellman's equation), we can characterize the equilibrium value function as

$$v^i(s;\pi_*) = \max_{\pi^i} \mathbb{E}_{a\sim(\pi^i,\pi_*^{-i}(s))} \left\{ r^i(s,a) + \gamma \sum_{\tilde{s}\in S} p(\tilde{s}|s,a) v^i(\tilde{s};\pi_*) \right\}.$$

- We define the Q-function, $Q^i(s,a;\pi_*)$, as the expression inside the "max and expectation",

$$Q^i(s,a;\pi_*) = r^i(s,a) + \gamma \sum_{\tilde{s}\in S} p(\tilde{s}|s,a) v^i(\tilde{s};\pi_*)$$

with $v^i(s;\pi_*) = \max_{\pi^i} \mathbb{E}_{a\sim(\pi^i,\pi_*^{-i}(s))} \{ Q^i(s,a;\pi_*) \}$.

FP for Stochastic Games

- We will consider a learning dynamic that combines fictitious play [Brown 49], [Robinson 51] with value function (or Q-function) iteration [Bertsekas 95]:
  - Players form beliefs on opponent strategies (using empirical frequencies and assuming opponent uses a stationary strategy).
  - Players also form beliefs on equilibrium value function, or Q-function.
  - Players choose a best response action in an "auxiliary game" given their beliefs (where the payoffs are given by the Q-function estimates).
- **The key challenge** is that the payoffs or value functions in these auxiliary games are non-stationary (unlike repeated play of stage games).

FP for Stochastic Games

- At stage $k \geq 0$, denote i's belief on $-i$'s strategy as π_k^{-i} and on her Q-function as Q_k^i and $Q_k^i(s, a^i, \pi_k^{-i}(s)) := \mathbb{E}_{a^{-i} \sim \pi_k^{-i}(s)}\{Q_k^i(s, a^i, a^{-i})\}$.
- Player i selects a best response $a_k^i(s)$ satisfying

$$\boxed{a_k^i(s) \in \arg\max_{a^i \in A^i} Q_k^i(s, a^i, \pi_k^{-i}(s))}.$$

- Player i updates her belief on player j's strategy as

$$\boxed{\pi_{k+1}^j(s) = \pi_k^j(s) + \alpha_k \Big(a_k^j(s) - \pi_k^j(s)\Big), \quad \text{for all } j \neq i \text{ and } s \in S}.$$

- Player i updates her belief on her Q-function as

$$\boxed{Q_{k+1}^i(s,a) = Q_k^i(s,a) + \beta_k \left(r^i(s,a) + \gamma \sum_{\tilde{s} \in S} p(\tilde{s}|s,a) v_k^i(\tilde{s}) - Q_k^i(s,a) \right)}$$

for all (s, a), with $v_k^i(\tilde{s}) = \max_{a^i \in A^i} Q_k^i(\tilde{s}, a^i, \pi_k^{-i}(\tilde{s}))$.

FP for Stochastic Games

- At stage $k \geq 0$, denote i's belief on $-i$'s strategy as π_k^{-i} and on her Q-function as Q_k^i and $Q_k^i(s, a^i, \pi_k^{-i}(s)) := \mathbb{E}_{a^{-i} \sim \pi_k^{-i}(s)}\{Q_k^i(s, a^i, a^{-i})\}$.
- Player i selects a best response $a_k^i(s)$ satisfying

$$\boxed{a_k^i(s) \in \arg\max_{a^i \in A^i} Q_k^i(s, a^i, \pi_k^{-i}(s))}.$$

- Player i updates her belief on player j's strategy as

$$\boxed{\pi_{k+1}^j(s) = \pi_k^j(s) + \alpha_k \Big(a_k^j(s) - \pi_k^j(s)\Big), \quad \text{for all } j \neq i \text{ and } s \in S}.$$

- Player i updates her belief on her Q-function as

$$\boxed{Q_{k+1}^i(s, a) = Q_k^i(s, a) + \beta_k \left(r^i(s, a) + \gamma \sum_{\tilde{s} \in S} p(\tilde{s}|s, a) v_k^i(\tilde{s}) - Q_k^i(s, a) \right)}$$

  for all (s, a), with $v_k^i(\tilde{s}) = \max_{a^i \in A^i} Q_k^i(\tilde{s}, a^i, \pi_k^{-i}(\tilde{s}))$.
- Reasonable & independent: the $\max_{a^i \in A^i}$ step is reasonable for the individual agent, but leads to local Q_k^i that differs among agents.

Two-timescale Learning Framework

- A key feature of our learning dynamics is that beliefs on Q-functions are updated at a slower timescale than beliefs on opponent strategies.
- This is consistent with the literature on evolutionary game theory [Ely and Yilankaya 01], [Sandholm 01] which postulate players' choices to be more dynamic than changes in their preferences.
    - Q-functions in auxiliary games can be viewed as slowly evolving player preferences.
- This assumption enables weakening the dependence between evolving strategies and Q-functions.
- We implement the two-timescale learning dynamics through the following assumption on the learning rates.

Assumption & Result

Assumption (Markov Chain)

Each state is visited infinitely often.

Holds if the stochastic game is irreducible: transition probabilities between any pair of states are positive for any joint action as in [Leslie et al. 21].

Assumption (Learning Rates)

(a) $\lim_{k\to\infty} \alpha_k = \lim_{k\to\infty} \beta_k = 0$ *and* $\sum_{k\geq 0} \alpha_k = \sum_{k\geq 0} \beta_k = \infty$.

(b) $\lim_{c\to\infty} \frac{\beta_k}{\alpha_k} = 0$.

Part (a) is classical in stochastic approximation theory.
Part (b) ensures two-timescale learning ($\beta_k \to 0$ faster than $\alpha_k \to 0$).

Theorem

Under these assumptions, for some stationary equilibrium (π_*^1, π_*^2) and the associated Q-function (Q_*^1, Q_*^2) of the zero-sum stochastic game, we have

$$(\pi_k^1, \pi_k^2) \to (\pi_*^1, \pi_*^2) \quad \textit{and} \quad (Q_k^1, Q_k^2) \to (Q_*^1, Q_*^2), \textit{w.p.1}, \quad \textit{as } k \to \infty.$$

Convergence Analysis

The evolution of the strategy and payoff estimates can be written as

$$\pi^i_{k+1}(s) = \pi^i_k(s) + \alpha_k(a^i_k(s) - \pi^i_k(s))$$
$$Q^i_{k+1}(s,a) = Q^i_k(s,a) + \beta_k\Big(r^i(s,a) + \gamma \sum_{\tilde{s}\in S} p(\tilde{s}|s,a) v^i_k(\tilde{s}) - Q^i_k(s,a)\Big)$$

for all (s,a), with $a^i_k(s) = \arg\max_{a^i} Q^i_k(s, a^i, \pi^{-i}_k(s))$ and $v^i_k(\tilde{s}) = \max_{a^i} Q^i_k(\tilde{s}, a^i, \pi^{-i}_k(\tilde{s}))$.

Two Challenges:

- Dynamics specific to an induced stage game is **coupled** with the dynamics at other stage games (due to $v^i_k(\tilde{s})$).
    - The two-timescale framework ($\beta_k/\alpha_k \to 0$) weakens this coupling.
- Each player updates Q^i using their local beliefs, induced stage games are **not necessarily zero-sum**.

Differential Inclusion Approximation

The discrete-time update can be written as

$$\pi_{k+1}^i(s) - \pi_k^i(s) \in \alpha_k \left(\arg\max_{a^i \in A^i} Q_k^i(s, a^i, \pi_k^{-i}(s)) - \pi_k^i(s) \right)$$
$$Q_{k+1}^i(s,a) - Q_k^i(s,a) = \alpha_k \; \varepsilon_k^i(s,a),$$

for each $i = 1, 2$, where the error term $\varepsilon_k(s,a) \approx \frac{\beta_k}{\alpha_k}$ is asymptotically negligible by the two-timescale assumption $\beta/\alpha \to 0$.

By the Differential Inclusion Approximation Theory [Benaim et al 05], we can approximate the update via

$$\dot{\pi}^i(s) \in \arg\max_{a^i \in A^i} Q^i(s, a^i, \pi^{-i}(s)) - \pi^i(s)$$
$$\dot{Q}^i(s,a) = 0,$$

for each $i = 1, 2$, which corresponds to the **continuous-time best response dynamics** of a game with stationary payoff functions $(Q^1(s,\cdot), Q^2(s,\cdot))$ since $\dot{Q}^i(s,a) = 0$.

Differential Inclusion Approximation

Lyapunov function

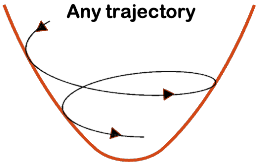
The Differential Inclusion Approximation Theory [Benaim et al 05] says that we can characterize the limit set of the discrete-time update via the differential inclusion (DI)

$$\dot{\pi}^i(s) \in \arg\max_{a^i \in A^i} Q^i(s, a^i, \pi^{-i}(s)) - \pi^i(s)$$

$$\dot{Q}^i(s, a) = 0$$

for each $i = 1, 2$ if we can find a Lyapunov function $V(\cdot)$. Particularly, we will have

$$\boxed{V(\pi_k(s), Q_k(s, \cdot)) \to 0}.$$



A Lyapunov Function

A continuous nonnegative function $V(\cdot)$:

- $V(x(t')) < V(x(t))$ for all $t' > t$ when $V(x(t)) > 0$
- $V(x(t')) = 0$ for all $t' > t$ when $V(x(t)) = 0$

for any solution $x(t)$ to the DI.

Lyapunov Function for Zero-sum Stochastic Games

- [Harris 98] showed that $V_H(\pi(s), Q(s,\cdot)) = \sum_i \max_{a^i \in A^i} \; Q^i(s, a^i, \pi^{-i}(s))$ is a **Lyapunov** function to the CT best response dynamics in a **zero-sum** game.
- Denote the best response of player i by $a^i_*(s)$. We have

$$\frac{d}{dt}\left(\max_{a^i \in A^i} \; Q^i(s, a^i, \pi^{-i}(s))\right) = Q^i(s, a^i_*(s), \dot{\pi}^{-i}(s)) \qquad \text{a.e.}$$

- Using $\dot{\pi}^{-i}(s) = a_*^{-i} - \pi^{-i}(s)$, we see V_H is **decreasing iff** non-negative $V_H > 0$:

$$\begin{aligned}\dot{V}_H &= \sum_i Q^i(s, a^i_*(s), a_*^{-i}(s)) - Q^i(s, a^i_*(s), \pi^{-i}(s)) \\ &= -V_H + \sum_i Q^i(s, a^i_*(s), a_*^{-i}(s)),\end{aligned}$$

  where the second term disappears since $Q^1(s,a) + Q^2(s,a) = 0$ **for all** a.
- Because of deviation from zero-sum structure in induced stage games, we develop a new Lyapunov function:

$$\boxed{V(\pi(s), Q(s,\cdot)) = \left(V_H(\pi(s), Q(s,\cdot)) - \lambda \max_a \left|\sum_i Q^i(s,a)\right|\right)_+}$$

  for any $\lambda \in (1, 1/\gamma)$.

Implications of the Lyapunov Function

- The new Lyapunov function and Differential Approximation Theory [Benaim et al 05] yield almost surely,

$$\boxed{V(\pi_k(s), Q_k(s,\cdot)) = \left(\sum_i \max_{a^i \in A^i} Q^i(s, a^i, \pi_k^{-i}(s)) - \lambda \max_a \left|\sum_i Q_k^i(s,a)\right|\right)_+ \to 0}$$

- This enables us to relate $\sum_i v_k^i(s) = \sum_i \max_{a^i \in A^i} Q^i(s, a^i, \pi_k^{-i}(s))$ with $\max_a |\sum_i Q_k^i(s,a)|$ and use stochastic approximation theory to show that the Q-function estimates are asymptotically zero sum:

$$\boxed{\lim_{k\to\infty} \max_a \left|\sum_i Q_k^i(s,a)\right| = 0}$$

  for each s and converge to equilibrium values.

- Since they track Shapley's minimax value iteration [Shapley 53], which converges to NE due to the γ-contracting property of the minimax VI operator.

Extensions - Model Free Learning

- In model-free learning, players do not know the transition probabilities and their own stage payoff function (only observe their realized stage payoffs).
- In this case, we use Q-learning, which is a stochastic form of value iteration [Watkins and Dayan 92].
- Without knowledge of transition probabilities, the players use the following estimate:

$$\sum_{\tilde{s}} p(\tilde{s}|s_k, a) v_k^i(\tilde{s}) \approx \hat{v}^i_{s_{k+1},k}$$

  if s_{k+1} is chosen with probability $p(s_{k+1}|s_k, a)$.
- Ensured by following the transitions of the Markov environment, making sample value of v at the successor state an unbiased estimate of the sum.
- Introduces additional stochastic approximation errors.
- Proper adjustment of learning rates within the two-time scale framework enables convergence to equilibrium values.

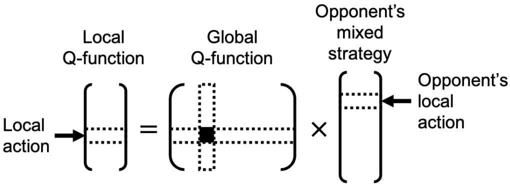
Extensions - Minimal Information

Also referred to as "Payoff-based" or "Radically Uncoupled" Learning

- Agents do not observe opponent's actions, therefore cannot form beliefs on opponent strategy.
- Instead, players estimate their local Q-function

$$q^i(s, a^i; \pi) := \mathbb{E}_{a^{-i} \sim \pi^{-i}(s)} \left\{ Q^i(s, a^i, a^{-i}; \pi) \right\}$$

based on the reward they receive since local Q-function carries information about opponent's strategy.



Minimal Information Case

- Players also form beliefs on the value function to estimate their continuation payoff:

$$v^i(s;\pi) = \max_{\pi^i} \ \mathbb{E}_{(a^i,a^{-i})\sim(\pi^i,\pi^{-i}(s))} \left\{ Q^i(s,a^i,a^{-i};\pi) \right\},$$

  which also captures the effect of their own strategy on their payoff.

- **Similar two-time scale learning framework:** Value functions updated at a slower timescale.
  - With adaptive learning rates, we can show asymptotic convergence to the equilibrium in two-player zero-sum stochastic games [Sayin∗, Zhang∗, Leslie, Başar, Ozdaglar, 21]

Minimal Information Case

- Players also form beliefs on the value function to estimate their continuation payoff:

$$v^i(s;\pi) = \max_{\pi^i} \ \mathbb{E}_{(a^i,a^{-i})\sim(\pi^i,\pi^{-i}(s))} \left\{ Q^i(s,a^i,a^{-i};\pi) \right\},$$

  which also captures the effect of their own strategy on their payoff.

- **Similar two-time scale learning framework:** Value functions updated at a slower timescale.
  - With adaptive learning rates, we can show asymptotic convergence to the equilibrium in two-player zero-sum stochastic games [Sayin*, Zhang*, Leslie, Başar, Ozdaglar, 21]

The results so far are all asymptotic:
Can we have non-asymptotic convergence rate (for best-response dynamics)?

Finite-Time Analysis for Minimal Information Case

- Limited results on rate analysis for best-response type learning in games.
  - Robinson's result $O(1/k^{\frac{1}{m+n-2}})$ (m, n sizes of actions sets) and Karlin's conjecture of $O(1/\sqrt{k})$ (proofs and disproofs for special cases [Daskalakis and Pan 14], [Abernethy, Lai and Wibisono 20])
  - Seminal result by [Harris 98] on rate of convergence of CT FP in zero-sum matrix games.
  - For stochastic games, all existing results for policy gradient or optimistic-gradient type methods.
- Our dynamics: Doubly smoothed best-response with value iteration:
  - Follows the two-timescale framework – change it to two-loop (see next slide) for finite-time analysis
  - Payoff-based and independent
  - No need to use adaptive stepsizes
- Sample complexity of $O(1/\epsilon)$ (to the Nash distribution) or $O(1/\epsilon^8)$ (to a Nash equilibrium) for matrix games and $O(1/\epsilon^8)$ (to a Nash equilibrium) for stochastic games.

Learning Dynamics (of Player i)

Inner Loop: Fix $\{\hat{v}^i_{s,t}\}_{s\in\mathcal{S}}$, and for $k = 0, 1, 2, \cdots, K-1$

- Given $\hat{q}^i_{s,t,k}$, player i updates $\hat{\pi}^i_{s,k}$ using doubly smoothed best-response:

$$\hat{\pi}^i_{s,t,k+1} = \underbrace{(1-\beta_k)\hat{\pi}^i_{s,t,k} + \beta_k \sigma^{\bar{\epsilon}}_{\tau}(\hat{q}^i_{s,t,k}),}_{\text{Taking a small (i.e., smooth) step towards the smoothed best-response}}$$

  where $\sigma^{\bar{\epsilon}}_{\tau}(q^i) := (1-\bar{\epsilon})\,\mathrm{argmax}_{\mu\in\Delta(\mathcal{A}^i)}\{\mu^\top q^i + \tau\cdot\nu(\mu)\} + \bar{\epsilon}\,\mathrm{Unif}(\mathcal{A}^i)$ is the smoothed best-response function with $\bar{\epsilon}$-perturbation, with $\nu(\mu)$ being the entropy of μ.

- Player i updates the local Q-function using temporal-difference learning:

$$\hat{q}^i_{s,t,k+1}(a^i) = \hat{q}^i_{s,t,k}(a^i) + \alpha_k \underbrace{\left(r^i_{t,k} + \gamma \hat{v}^i_{s_{k+1},t} - \hat{q}^i_{s,t,k}(a^i)\right)}_{\text{The temporal difference}}.$$

Note: To make TD-learning step work, we ensure policies evolve at a slower rate compared to that of q-functions (so that π_k is close to being *stationary*).

Learning Dynamics (of Player i)

Outer Loop: For $t = 1, \cdots, T-1$

- Player i updates the value function estimate $\{\hat{v}^i_{s,t}\}_{s \in \mathcal{S}}$ according to

$$\hat{v}^i_{s,t+1} = (\hat{\pi}^i_{s,t,K})^T \hat{q}^i_{s,t,K} \qquad \text{(An approximation of minimax VI)}$$

Note:

- $\hat{q}^i_{s,t,K}(a^i)$: local-Q function gives player i's expected payoff for action a^i.
- Player i computes expected payoff using the most recent strategy estimate $\hat{\pi}^i_{s,t,K}$.

Finite-Time Guarantees

Theorem

Under certain assumptions on stepsizes and $\bar{\epsilon} = \tau$, to achieve ϵ-approximate Nash equilibrium, the sample complexity is $\mathcal{O}(1/\epsilon^8)$.

Proof Sketch.

- Algorithm maintains 3 sets of coupled iterates $\{\hat{q}^i_{t,k}\}, \{\hat{v}^i_t\}, \{\hat{\pi}^i_{t,k}\}$.
- Construct Lyapunov functions for each.
- Challenge: Time-varying sampling policies due to "smooth best-response" $\Longrightarrow$ Time-inhomogeneous Markovian noise:
  - Establishing uniform ergodicity
  - An adaptive conditioning argument inspired by [Srikant and Ying, 2019]:

$$\mathbb{E}[\text{Update at } k] = \mathbb{E}[\mathbb{E}[\text{Update at } k \mid \mathcal{F}_{k-\text{mixing time}}]]$$

- Challenge: Highly-coupled iterates $\hat{q}^i_{s,t,k}$, $\hat{v}_{s,t}$, and $\hat{\pi}^i_{s,t,k}$
  - Establish Lyapunov drift inequalities for $\hat{q}^i_{s,t,k}$, $\hat{v}_{s,t}$, and $\hat{\pi}^i_{s,t,k}$
  - Solve the coupled Lyapunov inequalities to obtain the bound

□

Beyond Two-Player Games: Multi-Player Networked Markov Games

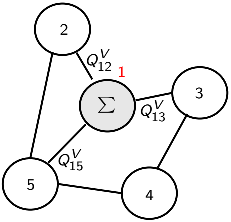
- All results presented so far are for two-player "zero sum" Markov games.
- Motivates a key question:

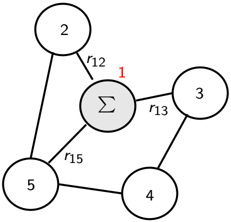
Are there other classes of stochastic games, beyond two-player zero-sum games, that allow tractable learning dynamics and equilibrium computation?

- Stochastic games with "Aligned Interests": [Sayin, Zhang, Ozdaglar 22] – Identical-interest Markov games with single controller.
- Networked Markov games – [Park, Zhang, Ozdaglar 23] .

Beyond Two-Player Games: Multi-Player Networked Markov Games

- Networked Markov Game (NMG) $(\mathcal{G} = (\mathcal{N}, \mathcal{E}_Q), \mathcal{S}, \mathcal{A}, \mathbb{P}, (r_i)_{i\in\mathcal{N}}, \gamma)$:
  - For any function $V : \mathcal{S} \to \mathbb{R}$ that defines
  $$Q_i^V(s,a) := r_i(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,a) V(s'),$$
  there exists a set of functions $(Q_{i,j}^V)_{(i,j)\in\mathcal{E}_Q}$ and a connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}_Q)$ such that $Q_i^V(s,a) = \sum_{j\in\mathcal{E}_{Q,i}} Q_{i,j}^V(s, a_i, a_j)$.
- Extends polymatrix (separable network) games [Bergman, Fokin 98] in normal form $(\mathcal{G} = (\mathcal{N}, \mathcal{E}), \mathcal{A}, (r_{i,j})_{(i,j)\in\mathcal{E}})$ where $r_i(a) = \sum_{\{j|(i,j)\in\mathcal{E}\}} r_{i,j}(a_i, a_j)$.





Characterization Results for NMG

Theorem (Sufficient and Necessary conditions for NMG)

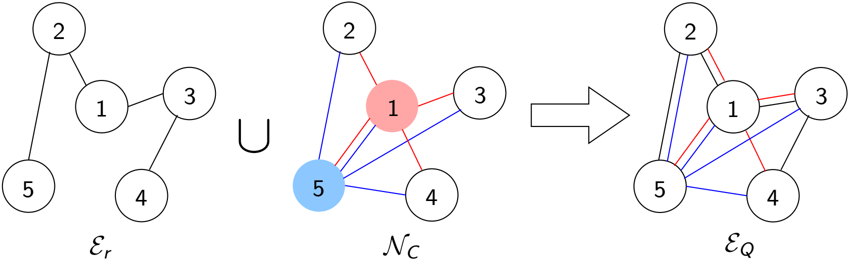
For a given graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}_Q)$, an MG $(\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathbb{P}, (r_i)_{i \in \mathcal{N}}, \gamma)$ is an NMG if and only if $r_i(s, a)$ and $\mathbb{P}(s'|s, \cdot)$ can be written as

$$r_i(s, a) = \sum_{j \in \mathcal{E}_{Q,i}} r_{ij}(s, a_i, a_j) \qquad \mathbb{P}(s'|s, a) = \sum_{j \in \mathcal{N}_C} w_j(s) \mathbb{P}_j(s'|s, a_j)$$

where the weights $w_j(s)$ satisfy $\sum_{j \in \mathcal{N}_C} w_j(s) = 1$ for all s, $\mathbb{P}_j$ is a probability distribution and $\mathcal{N}_C := \{i \mid (i, j) \in \mathcal{E}_Q \text{ for all } j \in \mathcal{N}\}$.

- Decomposable transition dynamics is an ensemble of transition dynamics controlled by single controllers:
  - For each $s \in \mathcal{S}$, sample $j \in \mathcal{N}_C$ with probability $w_j(s)$.
  - Then follow $\mathbb{P}_j(s'|s, a_j)$.
- Extends single-controller Markov games and turn-based Markov games.

Several results for NMG



Relationship between $\mathcal{E}_r$, $\mathcal{N}_C$, and $\mathcal{E}_Q$. The transition dynamics $\mathbb{P}$ is expressed as the ensemble of single controller $\mathcal{N}_C = \{1, 5\}$.

- An NMG is zero-sum if in addition $(\mathcal{G}, \mathcal{A}, (r_{i,j}(s))_{(i,j)\in\mathcal{E}_Q})$ is a zero-sum polymatrix game for all $s \in \mathcal{S}$.
- Paper shows fictitious play dynamics converge in NMGs when the underlying graph is a star network and hardness results for computing stationary NE and algorithms for computing nonstationary NE for non-star networks [Park, Zhang, Ozdaglar 23].

Conclusions

- We presented simple, reasonable and independent learning dynamics for stochastic games.
- For such dynamics, we present the first convergence guarantees to Nash equilibrium in zero-sum stochastic games.
- One key was two time-scale learning where estimates on opponent strategies are updated faster than estimates on value functions.
- Finite-sample analysis made possible following timescale-separation, but more delicate analysis of the coupled Lyapunov functions.

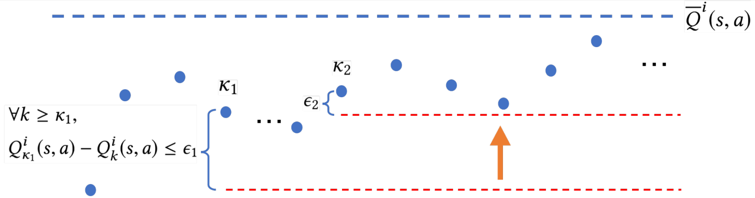
Ongoing and Future work:

- Convergence guarantees for potential stochastic games.
- Learning dynamics and non-asymptotic analysis for networked Markov games.
- Learning dynamics with function approximation to handle massively large state-action spaces.

Thank You!

Backup Slides

Identical-Interest Stochastic Games: Analysis (Cont'd)



- To this end, define a lower bound of $\Upsilon^i_k(s,a)$ as

$$\underline{u}^i_k := \min_{(s,a)} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \mathbb{E}_{a' \sim \pi_k(s)} \{ Q^i_k(s',a') \} - Q^i_k(s,a) \right\}$$

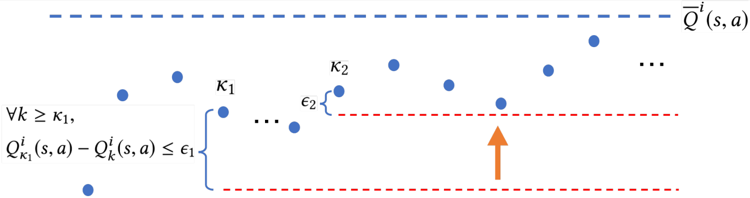
- One can show that $\underline{u}^i_k$ satisfies

$$\underline{u}^i_{k+1} \geq \underline{u}^i_k (1 - (1-\gamma)\beta_k) + \underline{e}_k,$$

with some absolutely summable sequence $\{\underline{e}_k\}$

- Unrolling it (using Gronwall Lemma), one can quantify the bound of $\underline{u}^i_k$ from below, and show $\liminf_{k_1 \to \infty} \inf_{k_2 \geq k_1} \sum_{k=k_1}^{k_2} \beta_k \underline{u}^i_k \geq 0$ (implies the desired result)

Identical-Interest Stochastic Games: Analysis (Cont'd)



- To this end, define a lower bound of $\Upsilon^i_k(s,a)$ as

$$\underline{u}^i_k := \min_{(s,a)} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)\mathbb{E}_{a' \sim \pi_k(s)}\{Q^i_k(s',a')\} - Q^i_k(s,a) \right\}$$

- One can show that $\underline{u}^i_k$ satisfies

$$\underline{u}^i_{k+1} \geq \underline{u}^i_k(1-(1-\gamma)\beta_k) + \underline{e}_k,$$

  with some absolutely summable sequence $\{\underline{e}_k\}$
- Unrolling it (using Gronwall Lemma), one can quantify the bound of $\underline{u}^i_k$ from below, and show $\liminf_{k_1 \to \infty} \inf_{k_2 \geq k_1} \sum_{k=k_1}^{k_2} \beta_k \underline{u}^i_k \geq 0$ (implies the desired result)
- Single-controller assumption is key to ensure the summability of $\{|\underline{e}_k|\}$