EE5180-Paper Presentation

Sidharthan SC(BE21B039) Amogh Kannan(AE21B004)

Indian Institute Of Technology, Madras

May 17, 2025

Unsynchronized decentralized Q learning: two timescale analysis by persistence

Bora Yongacoglu et. al, arXiv preprint, arXiv:2308.03239, 2023

Introduction- Multi Agent Reinforcement Learning

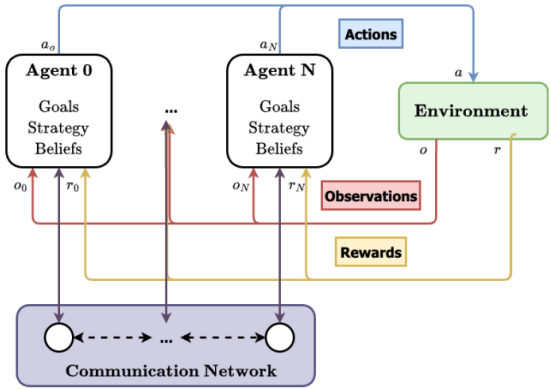

Figure 1: A visualization of a generalized multi-agent system following the iterative control process.

Introduction- Types of Games

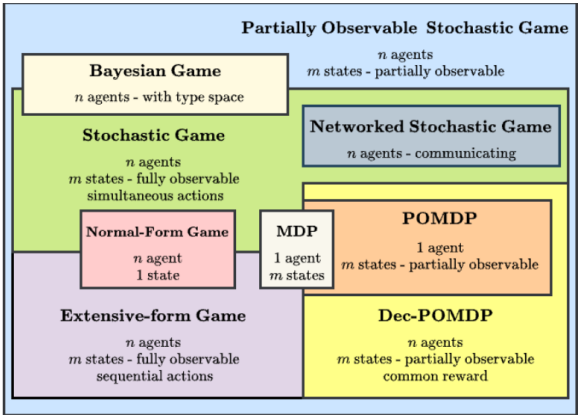

Figure 2: Models of Games: The overview of different models of multi-agent interactions is illustrated, from normal-form games to variations of stochastic games.

Introduction

- Systems composed of multiple autonomous agents interacting in a shared environment to achieve individual or collective goals.
- Non-Stationarity: The environment becomes dynamic as agents continuously adapt their policies based on interactions and feedback.
- Decentralized Information: Agents often operate with limited or incomplete information about the actions and policies of other agents.
- **Synchronized or Unsynchronized?**
- Most theoretical MARL frameworks avoid non-stationarity by synchronizing policy updates across agents. However, synchronization is impractical in decentralized or distributed systems.

Objective

- Introduce and analyze a decentralized Q-learning algorithm that works without synchronization, via inertial policy updating with persistent learning with non-decreasing learning rates.
- Prove that the method converges to equilibrium policies with high probability, overcoming non-stationarity.

Model: Stochastic Games in MARL

Stochastic Game:

$$G = (N, X, \{A_i, c_i, \gamma_i\}_{i \in N}, P, \nu_0)$$

- N: Set of agents
- X: Finite state space
- A_i: Agent i's finite action set; joint action set: $A = \prod_i A_i$
- $c_i : X \times A \to \mathbb{R}$: Cost function
- $\gamma_i \in [0, 1)$: Discount factor
- P: Transition kernel $P \in \mathcal{P}(X|X \times A)$
- ν_0: Initial state distribution

Game Dynamics & Information Structure

Time t progression:

- State x_t, joint action $a_t = (a_t^1, \ldots, a_t^N)$
- Agent i selects a_t^i using its policy
- Receives cost $c_i(x_t, a_t)$
- Transition: $x_{t+1} \sim P(\cdot|x_t, a_t)$

Local Action Learner (LAL) $\rightarrow$ Assumption 1:

- Agent i observes:
  - State x_t
  - Own action a_t^i
  - Own cost $c_i(x_t, a_t)$
  - Next state x_{t+1}
- Does **not** observe other agents' actions
- Observations are assumed noiseless

Policy Types and Stationarity

Policies:

- A policy $\pi^i = \{\pi_t^i\}_{t \geq 0}$ maps local history h_t^i to action distributions in A_i
- Π^i: Set of all policies for agent i

Stationary Policy:

$$x_t = x_k \implies \pi_t^i(a|h_t^i) = \pi^i(a|h_k^i)$$

- Stationary policies $\in \mathcal{P}(A_i|X)$
- Deterministic stationary policy: $\pi^i(x) = a^i$ with probability 1
- Π_{SD}^i: Set of deterministic stationary policies

Soft Policies and Best Responses

Soft Policy:

- ξ-soft: $\pi^i(a|x) \geq \xi$ for all x, a
- Soft policy: $\exists \xi > 0$ such that it is ξ-soft

Value Function:

$$J^i(\pi, \nu) := \mathbb{E}_\nu^\pi \left[\sum_{t=0}^{\infty} \gamma_i^t c_i(x_t, a_t) \right]$$

Best Response (with tolerance ϵ):

$$\pi^i \in \mathrm{BR}_\epsilon^i(\pi^{-i}) \iff J^i(\pi^i, \pi^{-i}) \leq \inf_{\tilde{\pi}^i} J^i(\tilde{\pi}^i, \pi^{-i}) + \epsilon$$

- ϵ-Equilibrium: $\pi = (\pi^1, \ldots, \pi^N)$ with each $\pi^i \in \mathrm{BR}_\epsilon^i(\pi^{-i})$
- Set of ϵ equilibrium stationary deterministic policies: $\Pi_{SD}^{\epsilon-eq}$.

Weakly Acyclic Stochastic Games

Strict Best Response Path:

- A sequence $\{\pi_k\}$ in Π_{SD} such that:

$$\pi^i_{k+1} \in \mathrm{BR}^i_0(\pi^{-i}_k), \quad \pi^i_{k+1} \neq \pi^i_k$$

  for a unique i at each step

Weakly Acyclic Game:

- $\Pi^{0\text{-eq}}_{SD} \neq \emptyset$
- Every joint policy has a path of best responses to an equilibrium

Examples:

- Stochastic teams: common cost function $c_i = c_j$
- Potential games with Markovian structure

Q-Functions in Stochastic Games

Q-function for Agent i:

$$Q^*_{\pi^{-i}}(x, a^i) = \mathbb{E}^{\pi^*}\left[\sum_{t=0}^{\infty} \gamma_i^t c_i(x_t, a_t) \,|\, x_0 = x, a_0^i = a^i\right]$$

where $\pi^* = (\pi^i, \pi^{-i})$ and $\pi^i \in \mathrm{BR}_0^i(\pi^{-i})$

ϵ-Greedy Policy w.r.t. Q^i:

$$\hat{\mathrm{BR}}^i_\epsilon(Q^i) = \left\{ \pi^{*i} \in \Pi^i_{SD} \;\middle|\; Q^i(x, \pi^{*i}(x)) \leq \min_{a \in A_i} Q^i(x, a) + \epsilon, \forall x \in X \right\}$$

Related Work: Q-Learning and Its Challenges

Single-Agent Q-Learning:

- Works well when environment is stationary
- Converges under:
  - Decaying learning rates (Proved by Tsitsiklis in [8])
  - Sufficient exploration (e.g., every state-action pair visited infinitely often)
- Decentralized MARL settings with stationary policies are sometimes treated as a set of isolated single agent RL problems

Boltzmann Action Selection (GLIE, or **G**reedy in the **L**imit with **I**nfinite **E**xploration):

- Common mechanism to balance exploration and exploitation; explores actions probabilistically using temperature parameter;temperature must decay slowly
- Singh et. al in [6] explore other GLIE algorithms.

In Multi-Agent Settings (MARL):

- Each agent faces a **non-stationary environment** due to learning by others
- No convergence guarantees unless other agents are stationary; Q-values may diverge
- The article David Leslie et. al ([2]) proves non-convergence.

Related Work: Algorithms for Local Action Learners

Local Action Learners (LAL): Each agent observes only the current state, own action and cost, and the state transition. The actions or costs of other agents cannot be observed.

In addition to previously mentioned problems in Q factor convergence, policy convergence also faces problems. Papers [3], etc. studied this. Proposed solutions include:

Multi-Timescale Learning:

- Separate learning rates:
  - Fast: Q-values or policy parameters
  - Slow: Policy updates or other parameters
- Mimics stationarity by decoupling adaptation
- Some agents may be made to learn at different timescales to others

Limitations:

- Many methods rely on:
  - Partial coordination
  - Shared update schedules
- Still challenged by full decentralization

Related Work: Regret Testers and Synchronization

Regret Testing Paradigm (Foster & Young), [4]:

- Similar concept to multi-time scale learning used; some iterates updated after every system interaction, others updated infrequently. The multi-time-scaling is done via update frequencies, not learning rates
- Good convergence guarantees proven for stateless repeated games
- Players periodically test whether switching policies reduces regret

Decentralized Q-Learning with Exploration Phases:

- Modification for the multi-state stochastic games
- Time is divided into **exploration phases** during which policies are kept fixed
- Learn Q-values, then update policy at end of phase-stationarity within each phase

Requirement: Synchronization

- All agents must agree on start/end of phases to ensure stationarity in each phase

Drawback:

- Coordination assumption is unrealistic in decentralized networks
- Communication delays and noise can desynchronize agents

Challenges in Prior Work vs. Contributions of This Paper

Challenges in Prior Work	Contributions of This Paper
Synchronized Learning Required: Agents must agree on fixed update times to ensure convergence.	**Decentralized Updates:** Each agent updates independently with no need for synchronization.
Non-Stationary Environment: Other agents adapt over time, making the environment non-stationary.	**Constant Learning Rates:** Quickly adapt to recent changes by discarding outdated information.
Instability Under Asynchronous Updates: Policy changes of one agent destabilize learning for others.	**Stability via Inertia:** Uses randomized inertia to slow changes and stabilize policy updates.
Requires Decaying Step Sizes: Limits adaptation speed in dynamic environments.	**Fast Adaptation:** Constant-rate Q-updates adapt faster in unsynchronized settings.
Lack of Robustness in Decentralized Scenarios: Noise and local decisions break convergence.	**Robust Design:** Each agent operates independently; no coordination or noise-sensitive steps.
No Theoretical Support for Unsynchronized Regret Testing: Prior regret testers rely on coordination.	**Formal Proof:** First to provide convergence bounds in fully unsynchronized MARL.
Limited to Specific Game Classes: Mostly restricted to weakly acyclic or symmetric games.	**Extended to General Games:** Supports ϵ-equilibria, team games, and cumber set convergence.
Delayed Error Correction: Decreasing rates slow correction of Q-value errors.	**Rapid Correction:** Persistent learning corrects faster in dynamic, real-time play.

Assumptions Overview

- Theoretical guarantees in unsynchronized MARL require careful assumptions.
- These relate to:
  - The transition dynamics of the stochastic game,
  - Agent-specific algorithmic parameters,
  - Independence of randomness in decentralized updates.

Assumption: Transition Kernel Reachability

Assumption 2 (Reachability)

For any pair of states $(s, s') \in \mathcal{X} \times \mathcal{X}$, there exists a sequence of joint actions $a'_0, \ldots, a'_H$ such that:

$$\Pr(x_{H+1} = s' \mid x_0 = s, a_0 = a'_0, \ldots, a_H = a'_H) > 0$$

Assumption: Transition Kernel Reachability

Assumption 2 (Reachability)

For any pair of states $(s, s') \in \mathcal{X} \times \mathcal{X}$, there exists a sequence of joint actions $a'_0, \ldots, a'_H$ such that:

$$\Pr(x_{H+1} = s' \mid x_0 = s, a_0 = a'_0, \ldots, a_H = a'_H) > 0$$

Interpretation: Any state can be reached from any other state using some finite joint action sequence. Ensures sufficient exploration is possible.

Assumption: Algorithmic Parameter Constraints

Assumption 3 (Tolerance and Exploration)

For each player $i \in N$:

- Suboptimality tolerance $\delta_i \in (0, \bar{\delta})$
- Exploration probability $\rho_i \in (0, \bar{\rho})$ with:

$$\| Q^{*j}_{\pi_{-j}} - Q^{*j}_{\hat{\pi}_{-j}} \|_\infty < \frac{1}{4} \min_i \min \left\{ \delta_i, \bar{\delta} - \delta_i \right\}$$

Assumption: Algorithmic Parameter Constraints

Assumption 3 (Tolerance and Exploration)

For each player $i \in N$:

- Suboptimality tolerance $\delta_i \in (0, \bar{\delta})$
- Exploration probability $\rho_i \in (0, \bar{\rho})$ with:

$$\| Q^{*j}_{\pi_{-j}} - Q^{*j}_{\hat{\pi}_{-j}} \|_\infty < \frac{1}{4} \min_i \min \left\{ \delta_i, \bar{\delta} - \delta_i \right\}$$

Key Insight: Controlled randomness and error thresholds ensure near-optimal Q-value estimation even in dynamic settings.

Assumption: Exploration Phase Bounds

Assumption 4 (Exploration Phase Lengths)

There exist integers $T, R \in \mathbb{N}$ such that:

$$T \leq T_k^i \leq RT \quad \forall i \in N, \forall k \geq 0$$

Assumption: Exploration Phase Bounds

Assumption 4 (Exploration Phase Lengths)

There exist integers $T, R \in \mathbb{N}$ such that:

$$T \leq T_k^i \leq RT \quad \forall i \in N, \forall k \geq 0$$

Implication: Ensures learning phases are neither too short (to prevent poor learning) nor too long (to avoid destabilization).

Assumption: Algorithmic Randomness Independence

Assumption 5 (Randomness Independence)

Define:

$$\mathcal{V}_1 = \bigcup_{i,t} \left\{ W_t, \tilde{\rho}_t^i, \tilde{u}_t^i, \tilde{\lambda}_t^i \right\}, \quad \mathcal{V}_2 = \bigcup_{i,t} \left\{ \tilde{\pi}_t^i(B) \mid B \subseteq \Pi_i^{\mathsf{SD}} \right\}$$

Then $\mathcal{V}_1 \cup \mathcal{V}_2$ are mutually independent.

Assumption: Algorithmic Randomness Independence

Assumption 5 (Randomness Independence)

Define:

$$\mathcal{V}_1 = \bigcup_{i,t} \left\{ W_t, \tilde{\rho}_t^i, \tilde{u}_t^i, \tilde{\lambda}_t^i \right\}, \quad \mathcal{V}_2 = \bigcup_{i,t} \left\{ \tilde{\pi}_t^i(B) \mid B \subseteq \Pi_i^{\mathrm{SD}} \right\}$$

Then $\mathcal{V}_1 \cup \mathcal{V}_2$ are mutually independent.

Purpose: Guarantees that random elements (actions, exploration, policy choice) across players are independent, enabling tractable analysis.

Proposed algorithm

Algorithm 1: Unsynchronized Decentralized Q-Learning

Set Parameters

$\{T_k^i\}_{k\geq 0}$: player i's sequence in $\mathbb{N}$ of learning phase lengths

Put $t_0^i = 0$ and $t_{k+1}^i = t_k^i + T_k^i$ for all $k \geq 0$.

$\rho^i \in (0,1)$: experimentation probability

$\lambda^i \in (0,1)$: inertia during policy update

$\delta^i \in (0,\infty)$: tolerance level for suboptimality

$\alpha^i \in (0,1)$: step-size parameter (also called the learning rate)

Initialize $\pi_0^i \in \Pi_{SD}^i$, $\widehat{Q}_0^i \in \mathbb{R}^{\mathbb{X}\times\mathbb{A}^i}$ (arbitrary)

for $k \geq 0$ (k^{th} exploration phase for agent i)

  for $t = t_k^i, t_k^i + 1, \ldots, t_{k+1}^i - 1$

    Observe x_t

    Select $a_t^i = \begin{cases} \pi_k^i(x_t), & \text{w.p. } 1-\rho^i \\ \bar{u}_t^i \sim \text{Unif}(\mathbb{A}^i), & \text{w.p. } \rho^i \end{cases}$

    Observe cost $c_t^i := c(x_t, \mathbf{a}_t)$, state x_{t+1}

    Put $\Delta_t^i = c_t^i + \gamma^i \min_{a^i} \widehat{Q}_t^i(x_{t+1}, a^i)$

    $\widehat{Q}_{t+1}^i(x_t, a_t^i) = (1-\alpha^i)\widehat{Q}_t^i(x_t, a_t^i) + \alpha^i \Delta_t^i$

    $\widehat{Q}_{t+1}^i(x, u^i) = \widehat{Q}_t^i(x, u^i)$, for all $(x, u^i) \neq (x_t, a_t^i)$

  if $\pi_k^i \in \widehat{\text{BR}}_{\delta^i}^i(\widehat{Q}_{t_{k+1}^i}^i)$, **then**

    $\pi_{k+1}^i \leftarrow \pi_k^i$

  else

    $\pi_{k+1}^i \leftarrow \begin{cases} \pi_k^i, & \text{w.p. } \lambda^i \\ \bar{\pi}_k^i \sim \text{Unif}\left(\widehat{\text{BR}}_{\delta^i}^i(\widehat{Q}_{t_{k+1}^i}^i)\right), & \text{w.p. } 1-\lambda^i \end{cases}$

Proof of convergence to equilibrium

Overview

Theorem

Let G be a weakly acyclic game and suppose each player $i \in N$ uses the given algorithm to play G. Suppose the assumptions of the previous slide hold. Let $\epsilon > 0$. There exists $\bar{\alpha}_\epsilon > 0$ and a function $\bar{T}_\epsilon \in (0,1)^N \times N \to N$ such that if:
$max_{i \in N} \alpha^i < \bar{\alpha}^\epsilon$, and $T > \bar{T}_\epsilon(\alpha, R)$, then:
$Pr(\phi_t \in \Pi_{SD}^{0-eq} \geq 1 - \epsilon)$ for sufficiently large t.

Proof of convergence to equilibrium

Overview

That is, if the relative frequency of policy update times of the players is controlled by some frequency R (players' update intervals lie between T and RT); and if learning rates are sufficiently small the policies converge to equilibrium with high probability.

Proof of convergence to equilibrium

Overview

Three major steps are involved.

- Introduce a set of equilibrium events, B_k, defined in terms of suitable time intervals: $[\tau_k^{min}, \tau_k^{max}]$. At each stage between the values of τ_k^{min} and τ_k^{max}, experience is accumulated by different agents in an unsynchronized fashion via a "baseline policy", and at the end of the learning period an update to policy is made if a better policy could be found. Define B_k such that over the interval τ_k^{min} to τ_k^{max}, the baseline policy does not change and was also an equilibrium.
- Argue that the probability of B_{k+1} given B_k can be lower bounded.
- Argue using the previous step that B_{k+L} given B_k can be lower bounded.

Proof of convergence to equilibrium

- Given historical data of the trajectory upto time t, a hypothetical Q factor $\hat{Q}$ can be defined, using random sampling of the various stochastic quantities (inertia, etc.) needed to propagate the trajectory, and assuming a frozen policy. It can be shown that this $\hat{Q}$ is bounded, using the standard recursive equation and the fact that discount factors γ are less than 1.
- It can also be shown that for sufficiently small learning rates and assuming no disruption caused by policy updates, the hypothetical $\hat{Q}$ can be brought arbitrarily close to the actual Q value (the agent can learn the true Q values using Q-learning)

Proof of convergence to equilibrium

- Updates due to other agents may destabilize the learning process. The terms τ_k^{min}, corresponding to, in the k^{th} "active phase", the first time at which an agent is able to update its policy (it may CHOOSE to not do so if its current policy seems optimal), and τ_k^{max}, the minimal time at which all agents will have had time to update policies at least once in epoch k, while retaining a buffer of T/N to the next update. The paper shows that such τ do exist, and furthermore, the maximum number of updates in this period between τ_k^{min} and τ_k^{max} is $R+1$.
- These results are used to prove that:
  - Given a history of states, with the last state being an equilibrium state, the probability that over the next active phase, τ_{k+1}^{min} to τ_{k+1}^{max}, the event B_{k+1} will be true, that is, the system will remain at equilibrium, can be lower bounded.
  - Using this result, the probability that after L more active phases, the system will still be at equilibrium, can also be lower bounded.

Simulation-Parameters

- 2-player game. $N = \{1, 2\}, X = \{s_0, s_1\}, A = \{a_0, a_1\}$.
- Discount factor $\gamma = 0.8$.
- Transition probability kernel:
  - $Pr(x_{t+1} = s_0 | x_t = s_0, a_t^1, a_t^2) = 0.5$
  - $Pr(x_{t+1} = s_0 | x_t = s_1, a_t^1 = a_t^2) = 0.25$
  - $Pr(x_{t+1} = s_0 | x_t = s_1, a_t^1 \neq a_t^2) = 0.9$
- The state dynamics of state s_0 do not depend on action taken, those of state s_1, however, do.

Simulation-Parameters

- Cost structure for state s_0:

	a_0	a_1
a_0	(0,0)	(2,2)
a_1	(2,2)	(0,0)

- Cost structure for state s_1:

	a_0	a_1
a_0	(10,10)	(11,11)
a_1	(11,11)	(10,10)

- Parameters were set as: $\rho^1 = \rho^2 = 0.05$, $\lambda^1 = \lambda^2 = 0.2$, $\delta^1 = \delta^2 = 0.5$, $\alpha^1 = \alpha^2 = 0.08$.

Simulation-Results

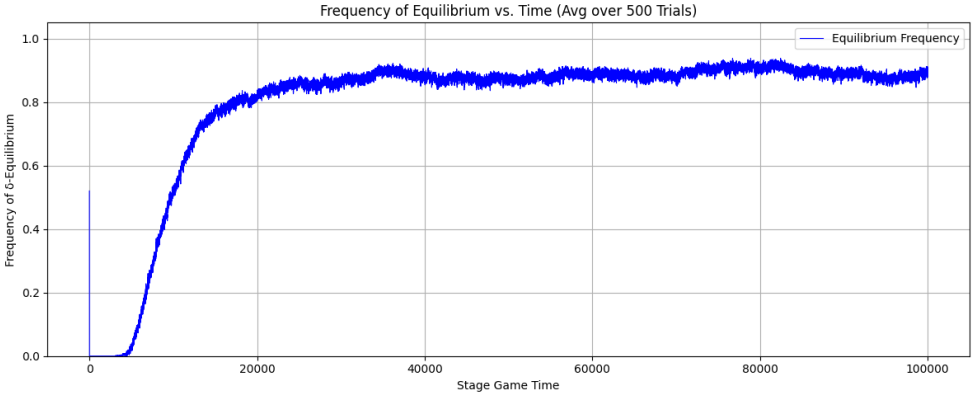

Figure 3: Frequency of equilibrium averaged over 500 Trials.

Simulation Results

Time (t)	Frequency
0	0.520
10000	0.518
20000	0.806
30000	0.854
40000	0.874
50000	0.862
75000	0.912
99999	0.882

Table 1: Observed frequency values at select simulation time points.

Possible extensions-Extending to Offline RL

- Offline data may not cover all state–action pairs uniformly.
- The paper relies on **Assumption 2**: full reachability and adequate exploration — which may not hold offline.
- Each agent would still learn independently from its own data slice — but now the data is fixed.
- Coordination must emerge **without communication** and **without new data** — this makes convergence analysis more difficult.

Contributions

Sidharthan S C Simulation, Literature review/survey, Extension of the paper
Amogh Kannan Proof and literature review/survey summary

References


S. V. Albrecht, F. Christianos, and L. Schäfer.
Multi-agent reinforcement learning: Foundations and modern approaches.
https://www.marl-article.com, 2024.


David Leslie et. al.
Individual q-learning in normal form games.
SIAM Journal on Control and Optimization, 44(2):495–514, 2005.


Dean Foster et. al.
On the nonconvergence of fictitious play in coordination games.
Games and Economic Behavior, 25(1):79–96, 1998.


Dean Foster et. al.
Regret testing: learning to play nash equilibrium without knowing you have an opponent.
Theoretical Economics, –(1):341–367, 2006.


Michael Bowling et. al.
Multiagent learning using a variable learning rate.

Artificial Intelligence, 136(2):215–250, 2002.


Satinder Singh et. al.
Convergence results for single-step on-policy reinforcement-learning algorithms.
Machine Learning, 38(–):287–308, 2000.


Ming Tan.
Multi-agent reinforcement learning independent vs cooperative agents.
In *ICML'93: Proceedings of the Tenth International Conference on International Conference on Machine Learning*, ICML, pages 330–337, –, 1993. ACM.


John N. Tsitsiklis.
Asynchronous stochastic approximation and q-learning.
Machine Learning, 16(–):185–202, 1994.