Finite-Sample Analysis of Decentralized Q-Learning for Stochastic Games

Zuguang Gao\textsuperscript{1}, Qianqian Ma\textsuperscript{2}, Tamer Başar\textsuperscript{3}, John R. Birge\textsuperscript{1}

\textsuperscript{1}The University of Chicago
\textsuperscript{2}LinkedIn Corporation
\textsuperscript{3}University of Illinois Urbana-Champaign
Overview

1. Introduction & Preliminaries
2. Algorithm & Sample Complexity
3. Numerical Experiment
4. Conclusion & Future Work
Introduction

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

- The frontier in many AI systems is now in multi-agent settings:

  (a) Distributed energy owners
  (b) Data center users
  (c) Supply chain participants

- Further advances critically depend on learning in multi-agent environments.
A (finite) discounted stochastic game has the following ingredients (Fink et. al., 1964).

- A finite set of agents, with the $i$-th agent referred to as agent $i$ for $i \in \{1, \ldots, N\} =: [N]$;
- A finite set $S$ of states;
- A finite set $\mathcal{A}^i$ of actions for each agent $i$;
- A nonnegative deterministic reward function $r^i$ for each agent $i$, which determines agent $i$'s reward, i.e., $r^i(s, a^1, \ldots, a^N) \in [0, r^i_{\text{max}}]$ at each state $s \in S$ and for each joint action $(a^1, \ldots, a^N) \in \mathcal{A} := \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$;
- A discount factor $\gamma^i \in (0, 1)$ for each agent $i$;
- A random initial state $s_0 \in S$;
- A transition kernel for the probability $P[s'|s, a^1, \ldots, a^N]$ of each state transition from $s \in S$ to $s' \in S$ for each joint $N$-tuple of actions $(a^1, \ldots, a^N) \in \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$. 
Policy, Value Function, and Equilibrium

- The set of deterministic stationary policies of agent $i$ is denoted by $\Pi^i := \{\pi^i : S \rightarrow A^i\}$.

- The objective of each agent $i$ is to find a policy $\pi^i \in \Pi^i$ that maximizes the total expected discounted reward, or the value function:

$$V^i(s) = \mathbb{E} \left[ \sum_{t=0}^{\infty} (\gamma^i)^t r^i(s_t, a_t) \left| s_0 = s \right. \right], \quad \forall s \in S.$$  \hspace{1cm} (1)

Markov perfect equilibrium

A joint policy $\pi^* = (\pi^*1, \ldots, \pi^*N) \in \Pi$ is a (Markov perfect) equilibrium if

$$V^i(\pi^*i, \pi^{*-i})(s) = \max_{\pi^i \in \Pi^i} V^i(\pi^i, \pi^{*-i})(s), \quad \forall s \in S, \ i \in \{1, \ldots, N\}.$$
Best Reply

• The Q-function (or action-value function) of agent $i$, $Q^i_\pi : S \times A^i \to \mathbb{R}$ of a joint policy $\pi$ is defined by

$$Q^i_\pi(s, a^i) = \mathbb{E}\left[\sum_{t=0}^{\infty} (\gamma^i)^t r^i(s_t, a^i_t, a^{-i}_t) \mid s_0 = s, a^i_0 = a^i\right], \quad \forall s \in S. \quad (2)$$

• Agent $i$’s set of deterministic best replies to $\pi^{-i}$ is

$$\Pi^{i}_{\pi^{-i}} = \left\{ \pi^{*i} \in \Pi^{i} : Q^{i}_{\pi^{-i}}(s, \pi^{*i}(s)) = \max_{a^{i} \in A^{i}} Q^{i}_{\pi^{-i}}(s, a^{i}), \quad \forall s \in S \right\}. $$

Strict best reply path

A sequence of deterministic joint policies $\pi_0, \pi_1, \ldots$ is called a strict best reply path if for each $k$, $\pi_k$ and $\pi_{k+1}$ differ for exactly one agent, say agent $i$, and $\pi^i_{k+1}$ is a strict best reply with respect to $\pi_k$. 
A discounted stochastic game is called weakly acyclic under strict best replies if there is a strict best reply path starting from each deterministic joint policy and ending at a deterministic equilibrium policy.

Best Reply Process with Inertia [Young 2004]

Set parameters
\( \lambda^i \in (0, 1) \): inertia

Initialize \( \pi_0^i \in \Pi^i \) (arbitrary)

Iterate \( k \geq 0 \) do
  
  if \( \pi_k^i \in \Pi^i_{\pi_{k-i}^i} \) then
    \( \pi_{k+1}^i = \pi_k^i \)
  
  else
    \( \pi_{k+1}^i = \begin{cases} 
    \pi_k^i & \text{w.p. } \lambda^i \\
    \text{any } \pi^i \in \Pi^i_{\pi_{k-i}^i} & \text{w.p. } (1 - \lambda^i)/|\Pi^i_{\pi_{k-i}^i}| 
  \end{cases} \)

end if

end
Best Reply Process with Inertia

BRPI Complexity

Let all agents update their deterministic policies according to the BRPI. We have that

\[ P \left[ \pi_k \in \Pi_{eq} \right] \geq 1 - \delta, \text{ provided that } k \geq \frac{L \cdot \log \delta}{\log \left( 1 - \left( \min_{j \in \{1, \ldots, N\}} \left\{ \frac{1 - \lambda_j}{|\Pi_j|} \cdot \lambda_j \right\} \right)^L \right)} + L. \]

Are we happy? Unfortunately, no!

BRPI assumes we know exactly each agent’s set of best replies \( \Pi_{\pi_k^i} \).

Use Q-learning to approximate the set \( \Pi_{\pi_k^i} \).

For each agent, the standard Q-learning is

\[
Q_{t+1}^i(s^t, a_t^i) = (1 - \eta^i_t)Q_t^i(s^t, a_t^i) + \eta^i_t \left[ r^i(s^t, a_t^i, a_t^{-i}) + \gamma^i \max_{a^i \in A^i} Q_t^i(s_{t+1}, a^i) \right],
\]

\[
Q_{t+1}^i(s, a^i) = Q_t(s, a^i), \quad \forall (s, a^i) \neq (s_t, a_t^i).
\]

Action selection: \( \epsilon \)-greedy.
Complete Algorithm [Arslan and Yüksel, 2016]

1: Initialize $\pi_0^i \in \Pi^i$ (arbitrary), $Q_0^i \in Q^i$ (arbitrary)
2: Receive $s_0$
3: for $k = 1, 2 \ldots$ do
4:   for $t = t_k, \ldots, t_{k+1} - 1$ do
5:     $a_t^i = \pi_k^i(s_t) := \begin{cases} 
    \pi_k^i(s_t), & \text{w.p. } 1 - \rho^i \\
    \text{any } a^i \in A^i, & \text{w.p. } \rho^i/|A^i| 
\end{cases}$
6:     Receive $r_t^i(s_t, a_t^i, a_t^{-i})$
7:     Receive $s_{t+1}$ (selected according to $P[ \cdot | s_t, a_t^i, a_t^{-i}]$)
8:     $Q_{t+1}^i(s_t, a_t^i) = (1 - \eta_t^i)Q_t^i(s_t, a_t^i) + \eta_t^i \left[ r_t^i(s_t, a_t^i, a_t^{-i}) + \gamma^i \max_{a^i \in A^i} Q_t^i(s_{t+1}, a^i) \right]$
9:     $Q_{t+1}^i(s, a^i) = Q_t^i(s, a^i)$, for all $(s, a^i) \neq (s_t, a_t^i)$
10: end for
11: $\Pi_{k+1}^i = \{ \hat{\pi}^i \in \Pi^i : Q_{k+1}^i(s, \hat{\pi}(s)) \geq \max_{a^i \in A^i} Q_{t_{k+1}}^i(s, a^i) - \frac{1}{2} \zeta^i, \text{ for all } s \}$
12: if $\pi_k^i \in \Pi_{k+1}^i$ then
13:    $\pi_{k+1}^i = \pi_k^i$
14: else
15:    $\pi_{k+1}^i = \begin{cases} 
    \pi_k^i, & \text{w.p. } \lambda^i \\
    \text{any } \pi^i \in \Pi_{k+1}^i, & \text{w.p. } (1 - \lambda^i)/|\Pi_{k+1}^i| 
\end{cases}$
16: end if
17: $Q_{t_{k+1}}^i \leftarrow$ projection of $Q_{t_{k+1}}^i$ onto $Q^i$
18: end for
### Sample Complexity Result

#### Assumption

There exist some $\kappa > 0$, and a finite integer $H \geq 1$, such that for any pair of states $(s', s)$, there exists a sequence of joint actions $\tilde{a}_0, \ldots, \tilde{a}_{H-1} \in \mathcal{A}$ such that

$$P[s_H = s' \mid (s_0, a_0, \ldots, a_{H-1}) = (s, \tilde{a}_0, \ldots, \tilde{a}_{H-1})] \geq \kappa.$$  

#### Main Theorem

Under certain assumptions, with carefully selected parameters $\eta^i_t, \rho^i, \zeta^i$, we have that for any $0 < \delta < 1$, $P[\pi_k \in \Pi_{eq}] \geq 1 - \delta$, provided that

$$T_k \geq \frac{c_0 A^{NH+1}}{\kappa \rho^{NH+1}} \left\{ \frac{1}{(1 - \bar{\gamma})^5 \epsilon^2} + \frac{(H + 1) \left( \log 4 \right) A^{NH} \kappa \rho^{NH} + 1}{1 - \bar{\gamma}} \right\},$$

$$K \geq \frac{\left( 1 - \tilde{\delta} \right)^2 \hat{\rho} - \tilde{\delta}^2}{\left[ \tilde{\delta} + (1 - \tilde{\delta}) \hat{\rho} \right]^2 \tilde{\delta}} L.$$
Numerical Experiment: A Grid-World Game

• We test the algorithm on the classical grid-world experiment [Sutton and Barto, 2018].
• Consider a $3 \times 3$ grid.

For each agent, there are 6 states where the agent can take two actions, and 3 states where the agent can take three actions. $|\Pi^i| = 2^6 \times 3^3 = 1728$. 

$$r^i(s, a^i, a^{-i}) = \begin{cases} 0, & s = (1,1) \\ -1, & s = \text{else} \end{cases}$$
Numerical Experiment: A Grid-World Game

- Obtain the set of all equilibrium joint policies.
- Check that the game is weakly acyclic
- Let $\rho^i = 0.4$, $\lambda^i = 0.3$, $\gamma^i = 0.75$, $\eta^i_k = 1/k^{0.5}$ for all $i$ and $k$.

**Figure:**

**a,** The fraction of times at which $\pi_k$ visits an equilibrium when $K$ ranges in the interval $[10, 1000]$.  
**b,** The fraction of times at which $\pi_k$ visits an equilibrium when $T$ ranges in the interval $[10, 2000]$.  

Conclusion & Future Work

Summary:

- We study the non-asymptotic convergence guarantee, namely, sample complexity, of the decentralized Q-learning algorithm in [Arslan and Yüksel, 2016].
- We apply linear function approximation to approximate the Q-functions in this general-sum stochastic game.
- We provide numerical experiments of both algorithms (with and without linear function approximation) on the classical Grid World game with minor modifications.

Future work:

- Further study on linear function approximation.
- Algorithms to converge to *optimal* equilibria
Thank you!

Questions? zuguang.gao@chicagobooth.edu