Decentralized Multi-Agent Reinforcement Learning for Continuous-Space Stochastic Games

Awni Altabaa[†], Bora Yongacoglu[‡], Serdar Yüksel[‡]

[†]Yale University, [‡] Queen's University

Overview

- Real-word sequential decision making problems often several challenges, including
  - complex environments with large/continuous state spaces,
  - multiple agents interacting with each other,
  - agents may not be able to communicate with each other.
- We propose a decentralized multi-agent reinforcement learning algorithm for continuous-space stochastic games.
- We characterize the stage-wise policy updating dynamics of the algorithm, as well as the global policy-updating dynamics of a broader class of best reply-based algorithms.

Goal

Our goal is to design an algorithm which is "rational" in a decentralized setting for each agent independently , and to analyze its global policy-updating dynamics (e.g., w.r.t. convergence to equilibria)

Outline

1. Background on Stochastic Games
2. Quantization of the state space
3. Decentralized Quantized Multi-Agent Q-Learning Algorithm
4. Analysis of Policy-Updating Dynamics

1. Background on Stochastic Games

2. Quantization of the state space

3. Decentralized Quantized Multi-Agent Q-Learning Algorithm

4. Analysis of Policy-Updating Dynamics

Stochastic Games

Definition (Stochastic game)

A stochastic game is a tuple $\left(\mathcal{N}, \mathbb{X}, \left\{\mathbb{U}^i\right\}_{i\in\mathcal{N}}, \mathcal{T}, \left\{c^i\right\}_{i\in\mathcal{N}}, \left\{\beta^i\right\}_{i\in\mathcal{N}}\right)$, where

1. $\mathcal{N} = \{1, ..., N\}$ is the set of $N > 1$ agents,
2. $\mathbb{X}$ is the state space, observed by all agents,
3. $\mathbb{U}^i$ is the action space of agent i; let $\mathbb{U} := \mathbb{U}^1 \times \cdots \times \mathbb{U}^N$,
4. $\mathcal{T} : \mathbb{X} \times \mathbb{U} \to \Delta(\mathbb{X})$ is the transition kernel, defining the probability of transitioning to $x' \in \mathbb{X}$ when the current state is $x \in \mathbb{X}$ and the agents take the joint action $\boldsymbol{u} \in \mathbb{U}$,
5. $c^i : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is the reward function of agent i giving the cost received when the system is in state x and the agents take joint action $\boldsymbol{u}$,
6. $\beta^i \in [0, 1)$ is the discount factor for each agent.

Objectives, best-replies and equilibrium

Each agent aims to minimize their own expected cumulative cost

$$J_x^i(\boldsymbol{\pi}) \coloneqq \mathbb{E}_x^{\boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} (\beta^i)^t c^i(X_t, \boldsymbol{U_t}) \right].$$

Definition (Best-reply)

Let $\epsilon \geq 0$ and let Γ^i be a subset of player i's policies. A policy $\pi^{*i} \in \Gamma^i$ is an ϵ-*best-reply* to $\boldsymbol{\pi}^{-i}$ in Γ^i if

$$J_x^i(\pi^{*i}, \boldsymbol{\pi}^{-i}) = \inf_{\pi^i \in \Gamma^i} J_x^i(\pi^i, \boldsymbol{\pi}^{-i}) + \epsilon, \ \forall x \in \mathbb{X}.$$

Furthermore, a 0-best-reply π^{*i} to $\boldsymbol{\pi}^{-i}$ is called a *strict best-reply* to $(\pi^i, \boldsymbol{\pi}^{-i})$ if $J_x^i(\pi^{*i}, \boldsymbol{\pi}^{-i}) < J_x^i(\pi^i, \boldsymbol{\pi}^{-i}), \quad$ for some $x \in \mathbb{X}$

Definition (Equilibrium)

For $\epsilon \geq 0$, a policy $\boldsymbol{\pi}^* \in \boldsymbol{\Gamma}$ is an ϵ-*equilibrium* in $\boldsymbol{\Gamma}$ if π^{*i} is an ϵ-best-reply to $\boldsymbol{\pi}^{*-i}$ for all $i = 1, ..., N$

1. Background on Stochastic Games

2. Quantization of the state space

3. Decentralized Quantized Multi-Agent Q-Learning Algorithm

4. Analysis of Policy-Updating Dynamics

Quantization of the state space

High-level idea: *group similar states into a finite set of representative bins and learn a value function on those bins.*

Consider the state space $\mathbb{X}$. Suppose $\mathbb{X}$ is a Borel subset of a Euclidean space.

Partition the state space $\mathbb{X}$ into M disjoint sets $\{B_i\}_{i=1}^{M}$, s.t. $\cup_i B_i = \mathbb{X}$. In each B_i, choose any representative state $y_i \in B_i$, and denote the quantized finite state space by $\mathbb{Y} = \{y_1, ..., y_M\}$.

We define the quantization mapping $q : \mathbb{X} \to \mathbb{Y}$ by $q(x) = y_i$ if $x \in B_i$.

This induces a "finite approximation MDP"

1. Background on Stochastic Games

2. Quantization of the state space

3. Decentralized Quantized Multi-Agent Q-Learning Algorithm

4. Analysis of Policy-Updating Dynamics

Decentralized Multi-Agent Q-Learning Algorithm

Algorithm 1: Algorithm for agent i

initialize $\pi_0^i \in \hat{\Pi}_q^i$, $Q_0^i \in \mathbb{Q}_q^i$ (arbitrary)

iterate $k \geq 0$ (`kth exploration phase`)

iterate $t = 1, ..., T_k$

Quantize state: $y_t^i = q^i(x_t)$

Choose action: $u_t^i = \begin{cases} \pi_k^i(y_t) & \text{w.p. } 1 - \rho^i \\ \text{any } u^i \in \mathbb{U}^i & \text{w.p. } \rho^i \end{cases}$

Receive $c^i(x_t, \boldsymbol{u}_t)$ and $x_{t+1} \sim \mathcal{T}\left(\cdot \,\middle|\, x_t, u_t^i, \boldsymbol{u}_t^{-i}\right)$

Quantize: $y_{t+1}^i = q^i(x_{t+1})$

$\alpha_t^i(y_t^i, u_t^i) = \left(1 + \sum_{s=t_k}^t \mathbb{I}\left\{(y_s^i, u_s^i) = (y_t^i, u_t^i)\right\}\right)^{-1}$

$\widehat{\mathsf{BR}}_\delta^i(Q_t^i) = \left\{\hat{\gamma} \in \hat{\Pi}_q^i \colon Q_t^i(y, \hat{\gamma}(y)) \leq \min_{u \in \mathbb{U}^i} Q_t^i(y, u) + \delta^i,\ \forall y \in \mathbb{Y}_q^i\right\}$

if $\pi_k^i \in \widehat{BR}_\delta^i(Q_t^i)$ **then** $\pi_{k+1}^i = \pi_k^i$;

else $\pi_{k+1}^i \in \widehat{\mathsf{BR}}_\delta^i(Q_t^i)$;

Reset Q_t^i to any $Q^i \in \mathbb{Q}_q^i$ (e.g.: $Q_t^i = 0$)

Near-optimality of policy updates

Theorem

Suppose all players use Algorithm 1 to select their actions. For any $\epsilon > 0$, *there exists* $\tilde{T}$ *such that* $T_k \geq \tilde{T}$ *implies*

$$\mathbb{P}\left[\left\|Q^i_{T_k} - \hat{Q}^{*i}_{\boldsymbol{\pi}^{-i}_{k,\rho}}\right\|_\infty < \epsilon\right] \geq 1 - \epsilon, \quad \forall k \geq 0, \tag{1}$$

where $\boldsymbol{\pi}_k$ *is the baseline joint policy during the* k^{th} *exploration phase and* $\boldsymbol{\pi}_{k,\rho}$ *is the perturbation of* $\boldsymbol{\pi}_k$ *that is used for action selection.*

Furthermore, for any $\eta > 0$, *there exists a fine-enough quantization* q^i *such that the greedy policy with respect to* $\hat{Q}^{*i}_{\boldsymbol{\pi}^{-i}_{\rho}}$ *is* η*-optimal in the MDP environment where the other agents act according to* $\boldsymbol{\pi}^{-i}_{\rho}$.

1. Background on Stochastic Games
2. Quantization of the state space
3. Decentralized Quantized Multi-Agent Q-Learning Algorithm
4. Analysis of Policy-Updating Dynamics

Generalized class of best reply-based algorithms

Consider the following class of algorithms for policy updates:

Algorithm 2: Generalized Policy-Updating Process (for agent i)

set parameters

$\psi^i(\cdot;\cdot)$ some prob. dist. over Π^i given a best-reply set

initialize $\pi_0^i \in \Pi^i$ (arbitrarily)

iterate $k \geq 0$

if $\pi_k^i \in BR^i(\boldsymbol{\pi}^{-i})$ **then** $\pi_{k+1}^i = \pi_k^i$;

else $\pi_{k+1}^i = \gamma^i \in \Pi^i$ w.p. $\psi^i(\gamma^i; \mathrm{BR}^i(\boldsymbol{\pi}^{-i}))$;

Analysis of policy-updating dynamics

Informal summary of results

The policy updating dynamics in the idealized process can be described as an *absorbing Markov chain* and a closed form expression can be derived for the probability of convergence to each equilibrium.

Stochastic algorithms which approximate such idealized policy updates can be shown to have the same dynamics in the limit.

Using these tools, the quantized continuous-space algorithm can be analyzed.