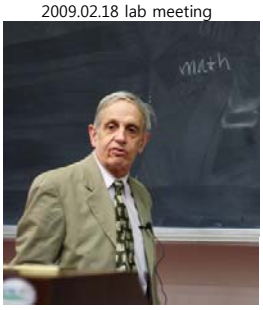
Nash Q-Learning

2009.02.18 lab meeting



GIST DCAS Lab, Han-Eol Kim

Index

Keyword : Nash Q-Learning

GIST, Department of Mechatronics

Markov Decision Process

Markov Decision Process game is a tuple $< S, A, r, p >$

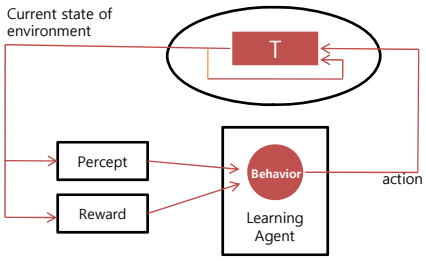
In a Markov decision process, the objective of the agent is to find a strategy (policy π) so as to maximize the expected sum of discounted rewards,

If it is optimal, solution of MDP, bellman equation can be described as

$$v(s, \pi^*) = \max_a \{r(s, a) + \beta \sum_{s'} p(s'|s, a) v(s', \pi^*)\}$$

S_0 is the initial state
r_t is the reward at time t
$0 \leq \beta < 1$ is the discount factor
π^* is the optimal strategy determined action by policy π

Q-Learning



At each time t, the agent chooses an action and observes it reward
The agent then updates it Q-values based on the following equation:
($0 \leq \alpha_t < 1$ is the learning rate)

$$Q_t(s_t, a_t) = (1 - \alpha_t) Q_t(s_t, a_t) + \alpha_t [r_t + \beta \max_b Q_t(s_{t+1}, a)]$$

Stochastic Games

In zero-sum game : $r^1(s,a^1,a^2)+r^2(s,a^1,a^2)=0$

In general-sum game : $r^1(s,a^1,a^2)+r^2(s,a^1,a^2)$ is not 0 or any constant

Nash Equilibrium

Definition of Nash Equilibrium

A Nash equilibrium is a joint strategy where each agent's is a best response to the others'. For a stochastic game, each agent's strategy is defined over the ent ire time horizon of the game.

In stochastic game, A Nash equilibrium point is a pair of strategies (π_*^1,π_*^2) such that for all $s \in S$

$$r^1(\pi_*^1,\pi_*^2) \geq r^1(\pi^1,\pi_*^2) \quad \text{for all} \quad \pi^1 \in A^1$$
$$r^2(\pi_*^1,\pi_*^2) \geq r^2(\pi_*^1,\pi^2) \quad \text{for all} \quad \pi^2 \in A^2$$

		agent1	
		a_1^1	a_1^2
agent2	a_2^1	(3,3)	(0,2)
	a_2^2	(2,0)	(1,1)

Example) Nash equilibrium point is (a_1^1,a_2^1)

Nash Q-function

Q-Learning Equation

$$Q_{t+1}(s_t, a_t) = (1 - \alpha_t)Q_t(s_t, a_t) + \alpha_t[r_t + \beta \max_b Q_t(s_{t+1}, a)]$$

Nash Q-Function

$$\text{Nash}Q_t^i(s') = \pi^1(s') \times \ldots \times \pi^n(s') \times Q_t^i(s')$$

Nash Q-Learning

$$Q_{t+1}^i(s, a^1, \ldots, a^n) = (1 - \alpha_t)Q_t^i(s, a^1, \ldots, a^n) + \alpha_t[\gamma_t^i + \beta \text{Nash}Q_t^i(s')]$$

Nash Q-Learning Algorithm

Initialize:
Let $t = 0$, get the initial state s_0.
Let the learning agent be indexed by i.
For all $s \in S$ and $a^j \in A^j$, $j = 1, \ldots, n$, let $Q_t^j(s, a^1, \ldots, a^n) = 0$.
Loop
Choose action a_t^i.
Observe $r_t^1, \ldots, r_t^n; a_t^1, \ldots, a_t^n$, and $s_{t+1} = s'$
Update Q_t^j for $j = 1, \ldots, n$
$Q_{t+1}^j(s, a^1, \ldots, a^n) = (1 - \alpha_t)Q_t^j(s, a^1, \ldots, a^n) + \alpha_t[r_t^j + \beta NashQ_t^j(s')]$
where $\alpha_t \in (0, 1)$ is the learning rate, and $NashQ_t^k(s')$
Let $t := t + 1$.

$$\text{Nash}Q_t^i(s') = \pi^1(s') \times \ldots \times \pi^n(s') \times Q_t^i(s')$$

Assumption

Assumption 1
Every state $s \in S$ and action $a^k \in A^k$ for $k = 1, \dots, n$ are visited infinitely often.

Assumption 2
The learning rate α_t satisfies the following conditions for all $s, t, a^1, \dots, a^n$:

1. $0 \leq \alpha_t(s,t,a^1,\dots,a^n) < 1, \ \sum_{t=0}^{\infty} \alpha_t(s,t,a^1,\dots,a^n) = \infty, \sum_{t=0}^{\infty} [\alpha_t(s,t,a^1,\dots,a^n)]^2 < \infty$
2. $\alpha_t(s,t,a^1,\dots,a^n) = 0$ if $(s,t,a^1,\dots,a^n) \neq (s_t, a_t^1, \dots, a_t^n)$

Assumption

Assumption 3
A Nash equilibrium $(\pi^1(s), \pi^2(s))$ for any 2-player stochastic game $(Q^1(s), Q^2(s))$ satisfies one of the following properties.

1. Nash equilibrium is global optimal

$$\forall \pi^1 \in A^1, \ \forall \pi^2 \in A^2 \qquad \pi_*^1(s) Q^k(s) \pi_*^2(s) \geq \pi^1(s) Q^k \pi^2(s)$$

2. If the Nash equilibrium is not a global optimal, then an agent receives a higher payoff when the other agent deviates from the Nash equilibrium strategy.

$$\forall \pi^1 \in A^1 \qquad \pi_*^1(s) Q^k(s) \pi^2(s) \geq \pi^1(s) Q^k \pi^2(s)$$

$$\forall \pi^2 \in A^2 \qquad \pi^1(s) Q^k(s) \pi_*^2(s) \geq \pi^1(s) Q^k \pi^2(s)$$

Convergence

Nash Q-Function

$$\mathrm{NashQ}_t^i(s') = \pi^1(s') \times \ldots \times \pi^n(s') \times Q_t^i(s')$$

Nash Q-Learning

$$\begin{aligned} Q_{t+1}^i(s, a^1, \ldots, a^n) &= (1-\alpha_t)Q_t^i(s, a^1, \ldots, a^n) + \alpha_t[\gamma_t^i + \beta \mathrm{NashQ}_t^i(s')] \\ &= (1-\alpha_t)Q_t^i(s, a^1, \ldots, a^n) + \alpha_t[\gamma_t^i + \beta\pi^1(s') \times \ldots \times \pi^n(s') \times Q_t^i(s')] \end{aligned}$$

Simplifying Nash-Q

$$Q_{t+1}^i(s) = (1-\alpha_t)Q_t^i(s) + \alpha_t[P_t^i Q^i(s')]$$

$$P_t^i Q^i(s') = \gamma_t^i + \beta\pi^1(s') \times \ldots \times \pi^n(s') \times Q_t^i(s')$$

Convergence Condition

$$||P_t Q - P_t Q_*|| \leq \beta ||Q - Q_*||$$

Convergence

Prove

$$\begin{aligned} ||P_t^i Q^i - P_t^i Q_*^i|| &= \gamma_t^i + \beta\pi^1 \times \ldots \times \pi^n \times Q^i - (\gamma_t^i + \beta\pi_*^1 \times \ldots \times \pi_*^n \times Q_*^i) \\ &= \beta(\pi^1 \times \ldots \times \pi^n \times Q^i - \pi_*^1 \times \ldots \times \pi_*^n \times Q_*^i) \end{aligned}$$

New notation $\sigma^i = \pi^i,\ \sigma^{-i} \equiv \sigma^1 \times \ldots \times \sigma^{i-1} \times \sigma^{i+1} \times \ldots \times \sigma^n$

$$\begin{aligned} \sigma^i\sigma^{-i}Q^i - \sigma_*^i\sigma_*^{-i}Q_*^i &\leq \sigma^i\sigma^{-i}Q^i - \sigma^i\sigma^{-i}Q_*^i \\ &= \sum \sigma^1 \times \ldots \times \sigma^n (Q^i - Q_*^i) \\ &\leq \sum \sigma^1 \times \ldots \times \sigma^n ||Q^i - Q_*^i|| = ||Q^i - Q_*^i|| \end{aligned}$$

$$\sum_{s'=1}^{m} p(s'|s, a^1, a^2) = 1$$

Finally, $||P_t^i Q^i - P_t^i Q_*^i|| \leq \beta ||Q^i - Q_*^i||$

THE END

Thanks for Listening