Markov Games and Multi-Objective Reinforcement Learning

Ather Gattami
Senior Research Scientist
RISE SICS
Stockholm, Sweden

February 28, 2018

Dynamical Systems

Let s_k, y_k, a_k be the state, observation, and action at time step k, respectively.

Deterministic model:

$$\begin{aligned} s_{k+1} &= f_k(s_k, a_k) \\ y_k &= g_k(s_k, a_k) \end{aligned}$$

Stochastic model (Markov Decision Process):

$$\begin{aligned} \mathbf{P}(s_{k+1} \mid s_k, a_k, s_{k-1}, a_{k-1}, ...) &= \mathbf{P}(s_{k+1} \mid s_k, a_k) \\ \mathbf{P}(y_k \mid s_k, a_k, s_{k-1}, a_{k-1}, ...) &= \mathbf{P}(y_k \mid s_k, a_k) \end{aligned}$$

We assume perfect state observation, that is $y_k = s_k$.

Dynamical Systems

Find the policy $a_k = \pi(s_k)$ that maximizes the average reward

$$V(s_0) = \mathbb{E}\left(\sum_{k=1}^{\infty} \delta^k r_k(s_k, a_k)\right)$$

where $0 < \delta < 1$ is the discount factor.

Stationary Bellman Equation

$$Q^*(s, a) = \mathbb{E}\left(r(s, a) + \delta \cdot Q^*(s_+, a_+)\right)$$

Bellman's Equation

Stationary Bellman Equation

$$Q^*(s,a) = \mathbb{E}\left(r(s,a) + \delta \cdot Q^*(s_+, a_+)\right)$$

$$\pi^\star(s) = \arg\max_\pi Q^\pi(s, \pi(s))$$

The optimal policy is

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

Finite Noncooperative Games

		Player 2	
		L	R
Player 1	U	$(1, 0)$	$(1, 3)$
	D	$(0, 2)$	$(2, 4)$

Bimatrix game (Q_1, Q_2) with payoffs

$$Q_1(a^1, a^2) = (a^1)^\intercal A_1 a^2, \qquad Q_2(a^1, a^2) = (a^1)^\intercal A_2 a^2$$

Optimal $a^1 = (0, 1)$, and $a^2 = (0, 1)$.

Finite Noncooperative Games

		Player 2	
		L	R
Player 1	U	$(4, 3)$	$(1, 1)$
	D	$(0, 0)$	$(3, 4)$

Bimatrix game (Q_1, Q_2) with payoffs

$$Q_1(a^1, a^2) = (a^1)^\intercal A_1 a^2, \qquad Q_2(a^1, a^2) = (a^1)^\intercal A_2 a^2$$

Mixed strategies: $a^1, a^2 \geq 0$, $a_1^1 + a_2^1 = 1$, $a_1^2 + a_2^2 = 1$.

Prisoner's Dilemma

Inefficient Nash Equilibrium:

		Prisoner 2	
		Cooperate	Defect
Prisoner 1	Cooperate	$(-1, -1)$	$(-10, 0)$
	Defect	$(0, -10)$	$(-5, -5)$

Stochastic Games

Given a Markov process (S, A_1, A_2, P) with

$$P(s, a^1, a^2, s_+) = \mathbf{P}(s_+ \mid s, a^1, a^2), \quad s \in S, \quad (a^1, a^2) \in A_1 \times A_2$$

and initial state $s_0 = s$.

Reward of Player 1:

$$V_1(s, \pi_1, \pi_2) = \mathbb{E}\left(\sum_{k=1}^{\infty} \delta^k r_k^1(s_k, \pi_1(s_k), \pi_2(s_k))\right)$$

Reward of Player 2:

$$V_2(s, \pi_1, \pi_2) = \mathbb{E}\left(\sum_{k=1}^{\infty} \delta^k r_k^2(s_k, \pi_1(s_k), \pi_2(s_k))\right)$$

Nash Equilibrium

Definition:

A Nash equilibrium is a pair of strategies $(\pi_1^\star, \pi_2^\star)$ such that for all $s \in S$,

$$V_1(s, \pi_1^\star, \pi_2^\star) \geq V_1(s, \pi_1, \pi_2^\star), \qquad \forall \pi_1$$

$$V_2(s, \pi_1^\star, \pi_2^\star) \geq V_2(s, \pi_1^\star, \pi_2), \qquad \forall \pi_2$$

Nash Equilibrium

Definition:

A Nash equilibrium is a pair of strategies $(\pi_1^\star, \pi_2^\star)$ such that for all $s \in S$,

$$V_1(s, \pi_1^\star, \pi_2^\star) \geq V_1(s, \pi_1, \pi_2^\star), \qquad \forall \pi_1$$

$$V_2(s, \pi_1^\star, \pi_2^\star) \geq V_2(s, \pi_1^\star, \pi_2), \qquad \forall \pi_2$$

Theorem (Filar and Vrieze, 1997)

Every (finite) stochastic game given by the tuple $(S, a^1, a^2, P, r_1, r_2)$ *possesses at least one Nash Equilibrium.*

Q-Learning (1 Player)

Let $s = s_k$ and $s_+ = s_{k+1}$.

Update rule with some $0 < \alpha_k(s_k, a_k) < 1$:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r(s,a) + \delta \max_{a_+} Q(s_+, a_+) - Q(s,a))$$

The optimal policy is estimated from $Q(s,a)$:

$$\pi(s) = \arg\max_a Q(s,a)$$

Q-Learning (1 Player)

Let $s = s_k$ and $s_+ = s_{k+1}$.

Update rule with some $0 < \alpha_k(s_k, a_k) < 1$:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r(s,a) + \delta \max_{a_+} Q(s_+, a_+) - Q(s,a))$$

The optimal policy is estimated from $Q(s,a)$:

$$\pi(s) = \arg\max_a Q(s,a)$$

Nash Equilibrium is **not equivalent** to maximizing the Q function!

Q-Learning (1 Player)

Theorem

Consider the Q-learning algorithm given by

$$Q(s,a) \leftarrow Q(s,a) + \alpha(s,a)(r(s,a) + \delta \max_{a_+} Q(s_+, a_+) - Q(s,a))$$

where $0 < \alpha_k(s_k, a_k) < 1$,

$$0 < \alpha_k(s,a) < 1, \qquad \sum_k \alpha_k(s,a) = \infty, \qquad \sum_k \alpha_k^2(s,a) < \infty, \qquad \forall\ (s,a)$$

The Q-learning algorithm converges to the optimal action-value function, $Q(s,a) \to Q^*(s,a)$.

Reinforcement Learning in Stochastic Games

Dynamic programming implies

$$Q_1^\star(s, a^1, a^2) = \mathbb{E}(r^1(s, a^1, a^2) + \delta Q_1^\star(s_+, a_+^1, a_+^2))$$

$$Q_2^\star(s, a^1, a^2) = \mathbb{E}(r^2(s, a^1, a^2) + \delta Q_2^\star(s_+, a_+^1, a_+^2))$$

Assumption

Assumption 1

The stochastic game satisfies one of the following properties:

(i) The Nash Equilibrium is global optimal.

$$V_1(s, \pi_1^\star, \pi_2^\star) \geq V_1(s, \pi_1, \pi_2), \qquad V_2(s, \pi_1^\star, \pi_2^\star) \geq V_2(s, \pi_1, \pi_2), \qquad \forall \pi_1, \pi_2$$

(ii) If the Nash Equilibrium is not global optimal, then an agent receives a higher payoff when the other agent deviates from the Nash Equilibrium strategy.

$$V_1(s, \pi_1^\star, \pi_2^\star) \leq V_1(s, \pi_1^\star, \pi_2), \qquad \forall \pi_2$$

$$V_2(s, \pi_1^\star, \pi_2^\star) \leq V_2(s, \pi_1, \pi_2^\star), \qquad \forall \pi_1$$

Reinforcement Learning in Stochastic Games

Theorem

Under Assumption 1, the Q-learning algorithm given by

$$Q_j(s,a) \leftarrow Q_j(s,a) + \alpha(s,a)(r^j(s,a) + \delta Q_j(s_+, \pi(s_+)) - Q_j(s,a)), \quad j = 1,2$$

where $\pi(s) = (\pi_1(s), \pi_2(s))$ is a pair of Nash Equilibrium strategies for the the bimatrix game (Q_1, Q_2)

$$0 < \alpha_k(s,a) < 1, \qquad \sum_k \alpha_k(s,a) = \infty, \qquad \sum_k \alpha_k^2(s,a) < \infty, \qquad \forall\ (s,a)$$

The Q-learning algorithm converges to the optimal action-value function, $Q(s,a) \to Q^*(s,a)$.

Zero-Sum Games

$$r^1(s, a^1, a^2) = -r^2(s, a^1, a^2)$$

Implies

$$V_1(s, \pi_1, \pi_2) = -V_2(s, \pi_1, \pi_2)$$

Assumption 1.(ii) is satisfied:

$$V_1(s, \pi_1^\star, \pi_2^\star) \leq V_1(s, \pi_1^\star, \pi_2), \qquad \forall \pi_2$$
$$V_2(s, \pi_1^\star, \pi_2^\star) \leq V_2(s, \pi_1, \pi_2^\star), \qquad \forall \pi_1$$

Nash Equilibrium in mixed strategies can be found by linear programming.

Q-Learning for Zero-Sum Games

Theorem (Q-learning for zero-sum games)

Consider the Q-learning algorithm given by

$$Q(s,a^1,a^2) \leftarrow Q(s,a^1,a^2) + \alpha(s,a^1,a^2)(r(s,a^1,a^2)+ \\ \delta \max_{a_+^1} \min_{a_+^2} Q(s_+,a_+^1,a_+^2) - Q(s,a^1,a^2))$$

where

$$0 < \alpha_k(s,a) < 1, \qquad \sum_k \alpha_k(s,a) = \infty, \qquad \sum_k \alpha_k^2(s,a) < \infty, \qquad \forall\ (s,a)$$

The Q-learning algorithm converges to the optimal action-value function, $Q(s,a) \to Q^*(s,a)$.

Multi-Objective Reinforcement Learning

- Industry: Production volume, cost, delivery, profit.
- Telecom: QoS, number of users, bit rate.
- Digital advertising: Reach vs. cost and return of investment.

Multi-Objective Reinforcement Learning

Single objective:

$$\mathbb{E}\left(\sum_{k=0}^{\infty}\delta^k r(s_k, a_k)\right) \geq \gamma$$

Maximize γ subject to the above inequality.

Multi-Objective Reinforcement Learning

Single objective:

$$\mathbb{E}\left(\sum_{k=0}^{\infty}\delta^k r(s_k,a_k)\right)\geq\gamma$$

Maximize γ subject to the above inequality.

Multiple objectives

$$\mathbb{E}\left(\sum_{k=0}^{\infty}\delta^k r^j(s_k,a_k)\right)\geq\gamma_j,\qquad j=0,...,J-1$$

Multi-Objective Reinforcement Learning

Lemma

Let $\beta_j = (1-\delta)\gamma_j$, *for* $j = 0, ..., J-1$. *If there exists a policy* $\pi \in \Pi$ *such that*

$$\mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k r^j(s_k, \pi(s_k))\right) \geq \gamma_j, \qquad j = 0, ..., J-1$$

then

$$\max_{\pi} \min_{j \in \mathbb{Z}_J} \mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k \left(r^j(s_k, \pi(s_k)) - \beta_j\right)\right) \geq 0$$

$$(1 + \delta + \delta^2 + \cdots)\beta_j = 1/(1-\delta) \cdot \beta_j = \gamma_j$$

Multi-Objective Reinforcement Learning

Lemma

Let $\beta_j = (1-\delta)\gamma_j$, *for* $j = 0, ..., J-1$. *If there exists a policy* π *such that*

$$\min_{j \in \mathbb{Z}_J} \mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k \left(r^j(s_k, \pi(s_k)) - \beta_j \right) \right) \geq 0$$

then

$$\mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k r^j(s_k, \pi(s_k)) \right) \geq \gamma_j, \qquad j = 0, ..., J-1$$

$$(1 + \delta + \delta^2 + \cdots)\beta_j = 1/(1-\delta) \cdot \beta_j = \gamma_j$$

Multi-Objective Reinforcement Learning

Theorem

Consider a Markov Decision Process given by (S, A, P) *and suppose that there exists a policy* π *such that*

$$\mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k r^j(s_k, \pi(s_k))\right) \geq \gamma_j, \qquad j = 0, ..., J-1$$

Let $\beta_j = (1-\delta)\gamma_j$ *and introduce*

$$r(s, a, j) \triangleq r^j(s, a) - \beta_j$$

Multi-Objective Reinforcement Learning

Theorem (Cont'd)

Let Q_k be given by the stochastic game Q-learning algorithm.

Then, $Q_k \to Q^\star$ as $k \to \infty$. Furthermore, the policy

$$\pi^\star(s) = \arg\max_{\pi} \min_{j \in \mathbb{Z}_J} \mathbb{E}\left(Q^\star(s, \pi(s), j)\right)$$

satisfies

$$\mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k r^j(s_k, \pi^\star(s_k))\right) \geq \gamma_j, \qquad j = 0, ..., J-1$$

Example

$$r^j(a) = \begin{cases} \frac{1}{2} & \text{if } a = j \\ 0 & \text{otherwise} \end{cases}$$

Let the discount factor be $\delta = \frac{1}{2}$ and let

$$\gamma_0 = \gamma_1 = \gamma_2 = \gamma = \frac{1}{3}$$

$$\mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k r^j(a_k)\right) \geq \frac{1}{3}, \qquad j = 0, 1, 2$$

Example

Now suppose that the agent takes action $a_k = 0$ with probability p_0. Then we have that

$$\mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k r^0(a_k)\right) = p_0$$

Similarly,

$$\mathbb{E}\left(\sum_{k=0}^{\infty} \delta^k r^j(a_k)\right) = p_j$$

for $j = 1, 2$.

Example

Suppose that $p_0 \leq p_1 \leq p_2$,

$$p_0 + p_1 + p_2 = 1$$

We have that

$$\frac{1}{3} = \frac{p_0 + p_1 + p_2}{3} \geq \sqrt[3]{p_0 p_1 p_2} \geq p_0$$

with equality if and only if $p_0 = p_1 = p_2 = \frac{1}{3}$.

The agent's mixed strategy is unique and given by $p_0 = p_1 = p_2 = \frac{1}{3}$.

Example

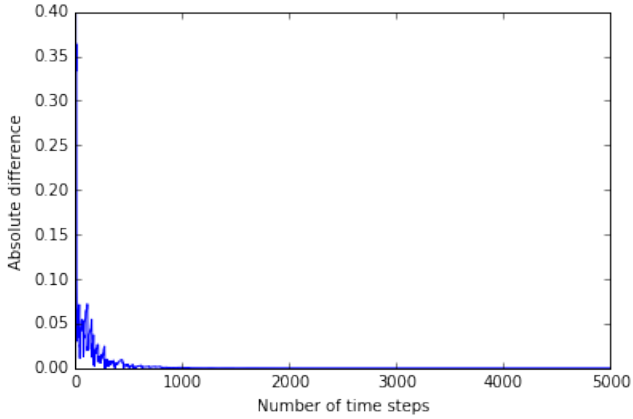

Figure: A plot of the maximum of $|p_0 - \hat{p}_0| + |p_1 - \hat{p}_1| + |p_2 - \hat{p}_2|$ over 1000 iterations, as a function of the number of time steps.

End of Presentation

QUESTIONS?