Markov Games

Jiří Vomlel

based on the papers

Michael L. Littman,
Markov Games as a framework for multiagent reinforcement learning
Proceedings of ICML'94, pages 157-163

Junling Hu and Michael P. Wellman,
Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm
Proceedings of ICML'98, pages 242-250

October 25, 2000

Matrices R^1 and R^2 of the rewards.

				deterministic str.	mixed str.
	0,0	+1,-1	-1,+1	yes	$p = 1/3$
	-1,+1	0,0	+1,-1	no	$p = 1/3$
	+1,-1	-1,+1	0, 0	no	$p = 1/3$

Optimal policy (strategy) for zero-sum games:

- **Deterministic**: $\pi \in A$, where A is a set of possible actions

$$a_\star = \arg \max_{a \in A^1} \min_{o \in A^2} a \cdot R(a, o) \cdot b$$

- **Probabilistic (mixed)**: $\pi \in PD(A)$, where $PD(A)$ is a set of discrete probability distribution over A

$$\pi^1_\star = \arg \max_{\pi^1 \in PD(A^1)} \min_{\pi^2 \in PD(A^2)} (\pi^1)^T R \, \pi^2$$

- Buyer with reservation price \$2 and seller with reservation price \$1.
- Transaction price is set to $(ask + bid)/2$ when $ask \leq bid$.

Ask\Bid	\$1	\$2	Equil Str 1	Equil Str 2
\$1	0, 1	0.5, 0.5	$\pi(a_1) = 1$	$\pi(a_1) = 0$
\$2	0, 0	1, 0	$\pi(a_2) = 0$	$\pi(a_2) = 1$
Equil Str 1	$\pi(b_1) = 1$	$\pi(b_2) = 0$		
Equil Str 2	$\pi(b_1) = 0$	$\pi(b_2) = 1$		

- General sum game: $0 + 1 = 0.5 + 0.5 = 1 + 0 \neq 0 + 0$.

A **pure strategy Nash equilibrium** $(a_\star, b_\star)$

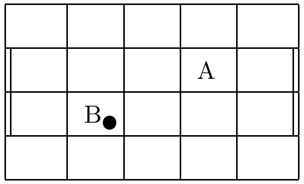
$$\begin{aligned} R^1(a_\star, b_\star) &\geq R^1(a, b_\star) \text{ for all } a \in A^1 \\ R^2(a_\star, b_\star) &\geq R^2(a_\star, b) \text{ for all } b \in A^2 \end{aligned}$$

A **mixed strategy Nash equilibrium** $(\pi^1_\star(a_1), \pi^1_\star(a_2)), (\pi^2_\star(b_1), \pi_\star(b_2))$

$$\begin{aligned} \pi^1_\star \cdot R^1 \cdot \pi^2_\star &\geq \pi^1 \cdot R^1 \cdot \pi^2_\star \quad \text{for all } \pi^1 \in PD(A^1) \\ \pi^1_\star \cdot R^2 \cdot \pi^2_\star &\geq \pi^1_\star \cdot R^2 \cdot \pi^2 \quad \text{for all } \pi^2 \in PD(A^2) \end{aligned}$$

There exists a mixed strategy equilibrium for any finite bimatrix game. (Nash,1951)

- Two players can move South, East, West, North or stand on the 4×5 grid.
- Once both players selected their actions the two moves are executed in random order
- When the player with the ball steps into appropriate goal, that player scores a point.
- When a player executes an action that would take it to the square occupied by the other player the ball goes to the stationary player and the move does not take place



- Set of states S
  - positions of player A and B, possession of the ball
- Collection of action sets for each agent $A^1, \ldots, A^k$
  - east, west, south, north, stand for agents A and B
- Transition function $S \times A^1 \times \ldots \times A^k \mapsto PD(S)$
  - $P(s_{t+1} \mid s_t, a, b)$
- Reward function for each agent $i = 1, \ldots, k$

$$R^i : S \times A^1 \times \ldots \times A^k \mapsto \Re$$

If s corresponds to agent i being with the ball in the goal then $R^i(s) = 1$.

GOAL: Each agent i attempts to find policy $\pi^i : S \mapsto PD(A^i)$ maximizing its *expected sum of discounted rewards.*

- Single stage (not repeated games) - *e.g. Prisoners dilemma*
- Multiple stages
  single state (iterated or repeated games) - *Rock, Scissors, Paper*
  multiple state (stochastic or Markov games) - *Soccer*

- One-player - *e.g. Bayesian Game Against Nature*
- Two-players
  zero-sum (matrix game), $R^2 = -R^1$ - *Rock, Scissors, Paper*
  general -sum (bimatrix game) $R^2 + R^1 \neq const.$ - *Bargaining*
- N-players

- Full information game
  (payoff matrices are known and actions can be observed)
- Incomplete information game
  (e.g. probabilistic information on opponents payoffs)

- Stationary strategy (does not depend on history)
  $\pi_t = \bar{\pi}$ for all t
- Markov strategy (depends on previous stage only)
  $\pi_t = f(s_{t-1}, a^1_{t-1}, a^2_{t-1})$
- Behavior strategy (depends on full history)
  $\pi_t = f(h_t)$, $h_t = (s_0, a^1_0, a^2_0, \ldots, s_{t-1}, a^1_{t-1}, a^2_{t-1})$

Every general-sum discounted stochastic game possesses at least one equilibrium point in stationary strategies. (Filar, Vrieze, 1997)

- Expected total rewards given policies π^1, π^2 and starting from state s *(value of state s)*:

$$V^1(s, \pi^1, \pi^2) = \sum_{t=0}^{\infty} \gamma^t \cdot E\left(r^1_t \mid \pi^1, \pi^2, s_0 = s\right)$$

$$V^2(s, \pi^1, \pi^2) = \sum_{t=0}^{\infty} \gamma^t \cdot E\left(r^2_t \mid \pi^1, \pi^2, s_0 = s\right)$$

- Expected rewards when taking actions a^1 and a^2 from state s and continuing optimally thereafter
  (quality of action a^1 against action a^2 in state s and vice-versa):

$$Q^1(s, a^1, a^2) = r^1(a^1, a^2, s) + \gamma \cdot \sum_{s'=1}^{N} P(s' \mid a^1, a^2, s) \cdot V^1(s', \pi^1_\star, \pi^2_\star)$$

$$Q^2(s, a^1, a^2) = r^2(a^1, a^2, s) + \gamma \cdot \sum_{s'=1}^{N} P(s' \mid a^1, a^2, s) \cdot V^2(s', \pi^1_\star, \pi^2_\star)$$

What does it mean "continuing optimally" in Zero-Sum Games?

In zero-sum games $Q^1(s, a^1, a^2) = -Q^2(s, a^1, a^2) = Q(s, a^1, a^2)$.

Littman (1994) suppose the opponent playing deterministic policy.
Hu and Wellman (1998) allowed mixed policy.

MiniMax policy:

$$\pi^1_\star(s) = \arg \max_{\pi^1 \in PD(A^1)} \min_{\pi^2 \in PD(A^2)} (\pi^1)^T Q(s)\, \pi^2$$

Value iteration algorithm repeats for $i = 1, 2, \ldots$ the following computation. For $s \in S, a^1 \in A^1, a^2 \in A^2$:

$$\begin{aligned} Q_i(s, a^1, a^2) &:= r^1(a^1, a^2, s) + \gamma \cdot \sum_{s'=1}^{N} P(s' \mid a^1, a^2, s) \cdot V^1_{i-1}(s') \\ V^1_i(s) &:= \max_{\pi^1(s) \in PD(A^1)} \min_{\pi^2(s) \in PD(A^2)} \pi^1(s)^T Q_i(s)\, \pi^2(s) \end{aligned}$$

What does it mean "continuing optimally" in General-Sum Games?

Nash equilibrium policy is proposed in (*Hu and Wellman, 1998*)

Value iteration algorithm repeats for $i = 1, 2, \ldots$ the following computation. For $s \in S, a^1 \in A^1, a^2 \in A^2$:

$$\begin{aligned} Q^1_i(s, a^1, a^2) &:= r^1(a^1, a^2, s) + \gamma \cdot \sum_{s'=1}^{N} P(s' \mid a^1, a^2, s) \cdot V^1_{i-1}(s') \\ Q^2_i(s, a^1, a^2) &:= r^2(a^1, a^2, s) + \gamma \cdot \sum_{s'=1}^{N} P(s' \mid a^1, a^2, s) \cdot V^2_{i-1}(s') \end{aligned}$$

For every $s \in S$ find a mixed strategy Nash equilibrium solution of the bimatrix game having matrices $Q^1_i(s)$ and $Q^2_i(s)$, so that we get policies (π^1_i, π^2_i). Then compute

$$\begin{aligned} V^1_i(s) &:= \pi^1_i(s)^T Q^1_i(s)\, \pi^2_i(s) \\ V^2_i(s) &:= \pi^1_i(s)^T Q^2_i(s)\, \pi^2_i(s) \end{aligned}$$

Q-learning

In the previous two formulations of Value Iteration:

- we had to solve the equations for every state $s \in S$.

Watkins (1989) proposed an alternative approach for MDP's (reinforcement learning):

- The values of V and Q are updated only for the observed state, where the agents got during the learning.
- Transition probability is not used.
- Single Agent Update:

$$Q_{i+1}(s, a) := (1 - \alpha_i) Q_i(s, a) + \alpha_i \left(r(s, a) + \gamma \max_{a' \in A} Q_i(s, a') \right)$$

Applicable to Markov games if:

- Agents can observe other agents' immediate payoffs and actions.
- Agents need not know other agents' payoff functions.
- Current state is known for sure.

Q-learning algorithm applied to Markov Games

(1) For $s \in S, a^1 \in A^1, a^2 \in A^2$ initialize $Q^1_0(s, a^1, a^2) = Q^2_0(s, a^1, a^2) = 1$.
(2) Set $s = s_0$.
(3) Find $\pi^1_0(s_0)$, a Nash equilibrium solution of the bimatrix game having matrices $Q^1_0(s_0)$ and $Q^2_0(s_0)$.
(4) For $i = 0, 1, 2, \ldots$ repeat:

- Choose action a^1_i based on the mixed strategy $\pi^1_i(s_i)$
- Observe a^2_i, r^1_i, r^2_i, and s_{i+1}.
- Find $\pi^1_i(s_{i+1}), \pi^2_i(s_{i+1})$, a Nash equilibrium solution of the bimatrix game having matrices $Q^1_i(s_{i+1})$ and $Q^2_i(s_{i+1})$
- Compute

$$\begin{aligned} V^1_{i+1}(s_{i+1}) &:= \pi^1_i(s_{i+1})^T Q^1_i(s_{i+1})\, \pi^2_i(s_{i+1}) \\ V^2_{i+1}(s_{i+1}) &:= \pi^1_i(s_{i+1})^T Q^2_i(s_{i+1})\, \pi^2_i(s_{i+1}) \\ Q^1_{i+1}(s_i, a^1, a^2) &:= (1 - \alpha_i) \cdot Q^1_i + \alpha_i \cdot \left(r^1_i + \gamma \cdot V^1_{i+1}(s_{i+1}) \right) \\ Q^2_{i+1}(s_i, a^1, a^2) &:= (1 - \alpha_i) \cdot Q^2_i + \alpha_i \cdot \left(r^2_i + \gamma \cdot V^2_{i+1}(s_{i+1}) \right) \end{aligned}$$

- *Every state and action have been visited infinitely often.*
- *Learning rate α_i satisfies:*

$$0 \leq \alpha_i < 1, \sum_{i=0}^{\infty} \alpha_i = \infty, \text{and} \sum_{i=0}^{\infty} \alpha_i^2 < \infty$$

- *Nash Equilibrium for any bimatrix game $Q_1(s), Q_2(s)$ is either global optimal or an agent receives higher payoff when the other agent deviates from the Nash equilibrium strategy*

Under the assumption listed above the multiagent Q-learning methods converges to the optimal (Nash equilibrium) values (for $k = 1, 2$):

$$Q^k(s, a^1, a^2) = r^k(a^1, a^2, s) + \gamma \cdot \sum_{s'=1}^{N} P(s' \mid a^1, a^2, s) \cdot V^k(s', \pi_\star^1, \pi_\star^2),$$

where $(\pi_\star^1, \pi_\star^2)$ is a Nash equilibrium solution for the stochastic game and $V^k(s', \pi_\star^1, \pi_\star^2)$ is expected total discounted reward of player k given $(\pi_\star^1, \pi_\star^2)$ when starting from state s'.

- Extending Partially Observed Markov Decision Processes (POMDP's) to games where instead of knowing a current state a probability distribution on the states is known. In that case Q-learning can not be directly applied.
- Replacing Nash Equilibrium used for selection of optimal policy by other criteria.
  - aggresive reservation price strategy
    Friedman (1991) for double auctions
  - strategy learned by gradient ascent
    Singh et all. (2000) for full information games