Cyclic Equilibria in Markov Games

Amy Greenwald

Brown University

with
Michael Littman and Martin Zinkevich

Games Research and Development

February 8, 2006

Agenda

Theorem

Value iteration converges to a stationary optimal policy in Markov decision processes.

Question

Does multiagent value iteration converge to a stationary equilibrium policy in Markov games?

Markov Decision Processes (MDPs)

Decision Process

- S is a set of states
- A is a set of actions
- $R : S \times A \to \mathbb{R}$ is a reward function
- $P[s_{t+1} \mid s_t, a_t, \ldots, s_0, a_0]$ is a probabilistic transition function that describes transitions between states, conditioned on past states and actions

MDP = Decision Process + Markov Property:

$$P[s_{t+1} \mid s_t, a_t, \ldots, s_0, a_0] = P[s_{t+1} \mid s_t, a_t]$$

$\forall t$, $\forall s_0, \ldots, s_t \in S$, $\forall a_0, \ldots, a_t \in A$

Bellman's Equations

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} P[s' \mid s,a] V^*(s') \tag{1}$$

$$V^*(s) = \max_{a \in A} Q^*(s,a) \tag{2}$$

Value Iteration

VI(MDP, γ)
- Inputs: discount factor γ
- Output: optimal state-value function V^*; optimal action-value function Q^*
- Initialize: V arbitrarily

REPEAT
  for all $s \in S$
    for all $a \in A$
      $Q(s,a) = R(s,a) + \gamma \sum_{s'} P[s' \mid s,a] V(s')$
    $V(s) = \max_a Q(s,a)$
FOREVER

Markov Games

Stochastic Game

- N is a set of players
- S is a set of states
- A_i is the ith player's set of actions
- $R_i(s, \vec{a})$ is the ith player's reward at state s given action vector $\vec{a}$
- $P[s_{t+1} \mid s_t, \vec{a}_t, \ldots, s_0, \vec{a}_0]$ is a probabilistic transition function that describes transitions between states, conditioned on past states and actions

Markov Game = Stochastic Game + Markov Property:

$$P[s_{t+1} \mid s_t, \vec{a}_t, \ldots, s_0, \vec{a}_0] = P[s_{t+1} \mid s_t, \vec{a}_t]$$

$\forall t,\ \forall s_0, \ldots, s_t \in S,\ \forall \vec{a}_0, \ldots, \vec{a}_t \in A$

Bellman's Analogue

$$Q_i^*(s, \vec{a}) = R_i(s, \vec{a}) + \gamma \sum_{s'} P[s' \mid s, \vec{a}] V_i^*(s') \tag{3}$$

$$V_i^*(s) = \sum_{\vec{a} \in A} \pi^*(s, \vec{a}) Q_i^*(s, \vec{a}) \tag{4}$$

Foe-VI $\pi^*(s) = (\sigma_1^*, \sigma_2^*)$, a minimax equilibrium policy
[Shapley 1953, Littman 1994]

Friend-VI $\pi^*(s) = e_{\vec{a}^*}$ where $\vec{a}^* \in \arg\max_{\vec{a} \in A} Q_i^*(s, \vec{a})$
[Littman 2001]

Nash-VI $\pi^*(s) \in \text{Nash}(Q_1^*(s), \ldots, Q_n^*(s))$
[Hu and Wellman 1998]

CE-VI $\pi^*(s) \in \text{CE}(Q_1^*(s), \ldots, Q_n^*(s))$
[G and Hall 2003]

Multiagent Value Iteration

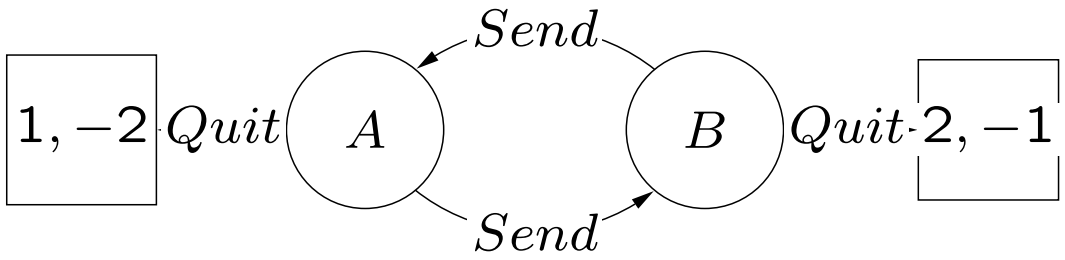
MULTI–VI(MGame, γ, f)
Inputs discount factor γ
selection mechanism f
Output equilibrium state-value function V^*
equilibrium action-value function Q^*
Initialize V arbitrarily

REPEAT
  for all $s \in S$
    for all $\vec{a} \in A$
      for all $i \in N$
        $Q_i(s, \vec{a}) = R_i(s, \vec{a}) + \gamma \sum_{s'} P[s' \mid s, \vec{a}] V_i(s')$
    $\pi(s) \in f(Q_1(s), \ldots, Q_n(s))$
    for all $i \in N$
      $V_i(s) = \sum_{\vec{a} \in A} \pi(s, \vec{a}) Q_i(s, \vec{a})$
FOREVER

Friend-or-Foe-VI *always* converges [Littman 2001]

Nash-VI and CE-VI converge to stationary equilibrium policies in zero-sum & common-interest Markov games [GZ and Hall 2005]

An Example



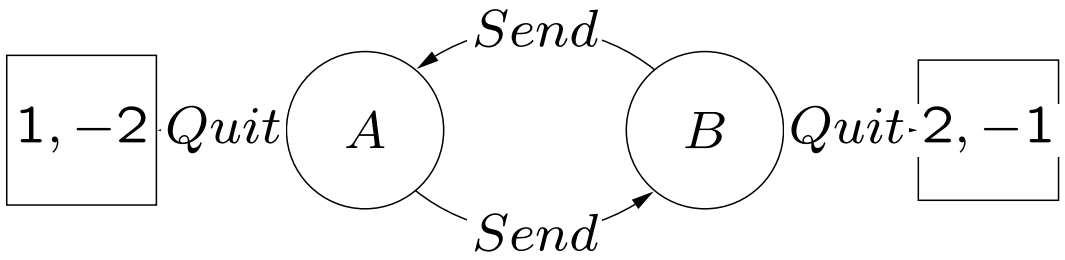
Observation

This game has no stationary deterministic equilibrium policy when $\gamma > \frac{1}{2}$.

Proof

(A quits, B quits)	$\Rightarrow$	A prefers send to quit ($2\gamma > 1$)
(A sends, B quits)	$\Rightarrow$	B prefers send to quit ($0 > -1$)
(A sends, B sends)	$\Rightarrow$	A prefers quit to send ($1 > 0$)
(A quits, B sends)	$\Rightarrow$	B prefers quit to send ($-1 > -2$)

A Cyclic Policy



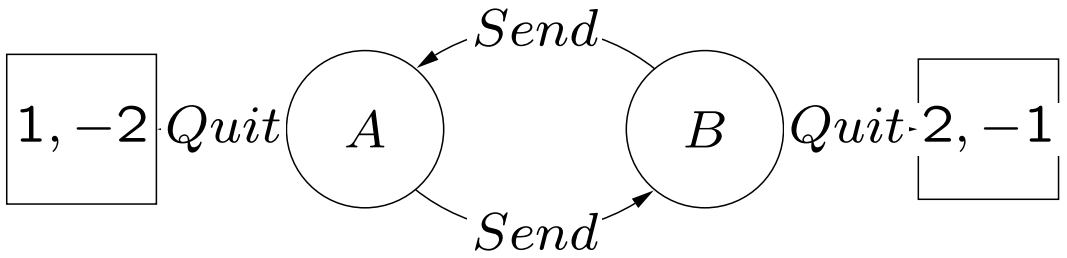
Observation

This game has a deterministic cyclic equilibrium policy when $\gamma = \frac{2}{3}$.

Example

	Policy	$V(A)$	$V(B)$
1	(A quits, B sends)	$(1, -2)$	$(\frac{8}{9}, -\frac{4}{9})$
2	(A sends, B sends)	$(\frac{4}{3}, -\frac{2}{3})$	$(\frac{8}{9}, -\frac{4}{9})$
3	(A sends, B quits)	$(\frac{4}{3}, -\frac{2}{3})$	$(2, -1)$
4	(A quits, B quits)	$(1, -2)$	$(2, -1)$

More Cyclic Policies



γ	(A sends, B sends)	Total
$\frac{2}{3}$	1	4
0.9	5	8
0.999999	693146	693149

Random Markov Games

$|N| = 2$

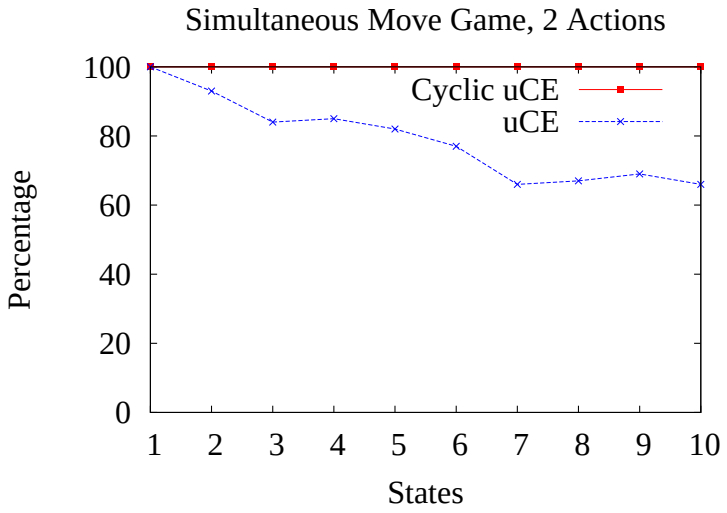
$|A| \in \{2, 3\}$

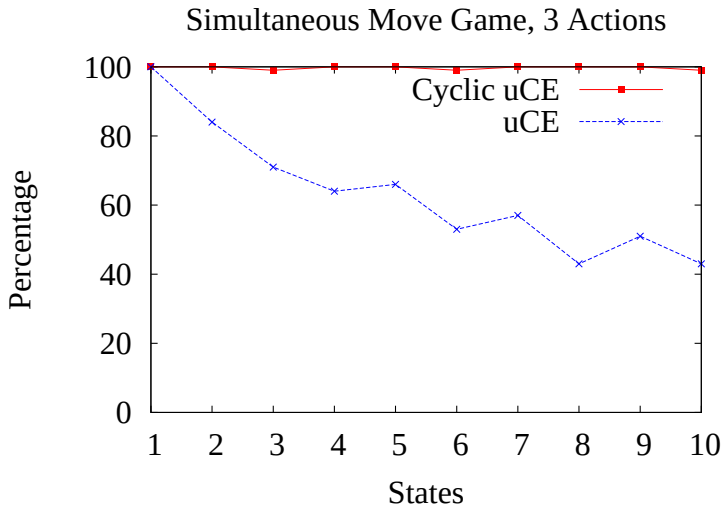
$|S| \in \{1, \ldots, 10\}$

Random Rewards $\in [0, 99]$

Random Deterministic Transitions

$\gamma = \frac{3}{4}$





Multiagent Q-Learning

Minimax-Q Learning [Littman 1994]

- provably converges to stationary minimax equilibrium policies in zero-sum Markov games

Nash-Q Learning [Hu and Wellman 1998]
Correlated-Q Learning [G and Hall 2003]

- converge empiricially to stationary equilibrium policies on a testbed of general-sum Markov games