Correlated-Q Learning and Cyclic Equilibria in Markov games

Haoqi Zhang
Correlated-Q Learning

Greeenwald and Hall (2003)

• Setting: general sum Markov games
• Goal: convergence (reach equilibrium), payoff
• Means: CE-Q
• Results: empirical convergence in experiments
• Assumptions: observable reward, umpire for CE selection
• Strong? Weak? What do you think?
Markov Games

- State transitions only dependent on current state and action
- Q-values over states and action-vectors over agents
- Don’t always exist deterministic actions that maximize each agent’s rewards
- Each agent plays an action profile with a certain probability
Q-values

- Use Q values to find best action (in single player, argmax a..)
- In Markov game, can use Nash-Q, CE-Q, ..., which use Q-values as the entries to a stage game and compute the equilibria.
- Play according to probabilities in the optimal strategy (your own part)
Nash equilibrium vs. Correlated equilibrium

Nash Eq.
- vector of independent probability distributions over actions
- No unilateral deviation given everyone else is playing the equilibrium

Correlated Eq.
- joint probability distribution (e.g. traffic light)
- No unilateral deviation given that others believe you are playing the equilibrium
Why CE?

- Easily computable with linear programming
- Higher rewards than Nash Equilibrium
- No-regret algorithms converge to CE (Foster and Vohra)
- Actions chosen independently (but based on commonly observed private signal)
LP to solve for CE

<table>
<thead>
<tr>
<th></th>
<th>$L$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>6,6</td>
<td>2,7</td>
</tr>
<tr>
<td>$B$</td>
<td>7,2</td>
<td>0,0</td>
</tr>
</tbody>
</table>

The correlated equilibria in this game are described by the probability constraints $\pi_{TL} + \pi_{TR} + \pi_{BL} + \pi_{BR} = 1$ and $\pi_{TL}, \pi_{TR}, \pi_{BL}, \pi_{BR} \geq 0$ together with the following so-called "rationality" constraints:

$-1\pi_{TL} + 2\pi_{TR} \geq 0$

$1\pi_{BL} - 2\pi_{BR} \geq 0$

$-1\pi_{TL} + 2\pi_{BL} \geq 0$

$1\pi_{TR} - 2\pi_{BR} \geq 0$

These are constraints, need an objective
Multiple Equilibria

• There are many equilibria (can be much more than Nash!)
• Need a way to break ties
• Can ensure equilibrium value is the same (although maybe not equilibrium policy)
• 4 variants
  – Maximize the sum of players’ rewards (uCE-Q)
  – Maximin of the players rewards (eCE-Q)
  – Maximax of the players rewards (rCE-Q)
  – Maximize the maximum of each individual player (lCE-Q)
Experiments

3 grid games
- Exists both deterministic and nondeterministic equilibrium
- Q-values converged (in 500000+ iterations)
- \{u,e,r\}CE-Q with best score performance (discount factor of 0.9)

Soccer game
- Zero-sum, no deterministic eq.
- uCE (and others) still converges
Where are we?

- Some positive results, but highly enforced coordination
- Problem: multiplicity of equilibria
- Are these results useful for anything? Why should we care?
Cyclic Equilibria in Markov Games

Zinkervich, Greenwald, Littman

• Setting: General sum Markov games
• Negative result: Q-values alone is insufficient to guarantee convergence
• Positive result: Can often get to cyclic equilibrium
• Assumptions: offline (what happened to learning?)

• How do we interpret these results? Why should we care?
Policy

- Stationary policy - set distribution for state, action vector pairs
- Non-stationary policy - a sequence of policies played at each iteration
- Cyclic policy - a non-stationary policy that is cyclic
Policy

• Stationary policy - set distribution for state, action vector pairs
• Non-stationary policy - a sequence of policies played at each iteration
• Cyclic policy - a non-stationary policy that is cyclic
Figure 1: An example of a NoSDE game. Here, $S = \{1, 2\}$, $A_{1,1} = A_{2,2} = \{keep, send\}$, $A_{1,2} = A_{2,1} = \{noop\}$, $T(1, \text{keep, noop}) = 1$, $T(1, \text{send, noop}) = 2$, $T(2, \text{noop, keep}) = 2$, $T(2, \text{noop, send}) = 1$, and $\gamma = 3/4$. In the unique stationary equilibrium, Player 1 sends with probability $2/3$ and Player 2 sends with probability $5/12$. 
NoSDE game (nasty)

• Turn-taking game
• No deterministic stationary policy
• Every NoSDE game has a unique nondeterministic stationary equilibrium policy
• Negative result
  For any NoSDE game, there exists another NoSDE game (differing in only rewards) with its own stationary policy such that the Q values are equal but the policies are different and the values are different.
• How do we interpret this?
Cyclic Equilibria

• Cyclic correlated equilibrium: a cyclic policy that is a correlated equilibrium
• CE: for any round in the cycle, playing based on observed signal has higher value (based on Q’s) than deviating.
• Can use value iteration to derive cyclic CE
Value Iteration

2. Use $V$'s from last iteration to update current $Q$'s
3. Compute policy using $f(Q)$
4. Update current $V$'s using current $Q$'s
GetCycle

1. Run value iteration
2. Find minimal distance between final round $V_T$ and any other round (that is less than maxCycles away), where distance is max difference between any state
3. Set the policies to the policies between these two rounds
Fact 1 If \( d(V^T, V^{T-1}) = \epsilon \) in \textit{GetStrategy}, then \textit{GetStrategy} returns an \( \frac{\epsilon \gamma}{1-\gamma} \)-correlated equilibrium.

Fact 2 If \textit{GetCycle} returns a cyclic policy of length \( k \) and \( d(V^T, V^{T-k}) = \epsilon \), then \textit{GetCycle} returns an \( \frac{\epsilon \gamma}{1-\gamma^k} \)-correlated cyclic equilibrium.
Theorems

Theorem 2: Given selection rule uCE, for every NoSDE game, there exists a cyclic CE.

Theorem 3: Given selection rule uCE, for any NoSDE game, ValueIteration does not converge to the optimal stationary policy.

Theorem 4: Given the game in Figure 1, no equilibrium selection rule f converges to the optimal stationary policy.

Strong? Weak? Which one?
Experiments

• Check convergence by running metric:

\[
\max_{t \in \{1, \ldots, k\}} d_{\Gamma}(V^{1001-t}, V^{1001-(t+k)}) \leq 0.0001,
\]

Check if deterministic equilibria exist by enumerating over every deterministic policy and running policy evaluation for 1000 iterations to estimate $V$ and $Q$. 
Results

Test on turn based game and small simultaneous games, reached Cyclic CE with uCE almost always. With 10 states and 3 actions in simultaneous games, no techniques converged
What does this all mean?

• How negative are the results?
• How do we feel about all the assumptions?
• What are the positive results? Are they useful? Why are cyclic equilibria interesting?

• What about policy iteration?
The End :)