Markov Games Framework and Results about Cooperative Multiagent Systems

Presented by Chen Tang

June 8, 2022

1. From Matrix Games and MDP to Markov Games

2. RL in Cooperative Multiagent Systems

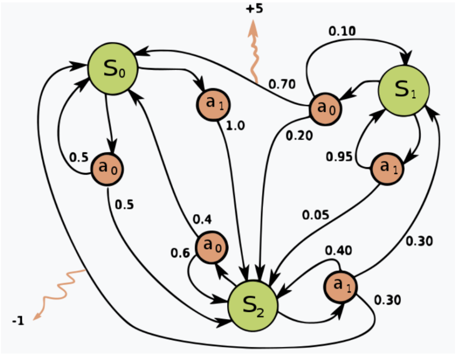
1. From Matrix Games and MDP to Markov Games

2. RL in Cooperative Multiagent Systems

Markov games as a framework for multi-agent reinforcement learning
-*Machine learning proceedings* (1994)

Michael L. Littman, the computer science professor at Brown University, studying machine learning and working to engage broadly about applications and implications of artificial intelligence.

Matrix Games and Markov Decision Process



Matrix games (two-player zero-sum)

$$V = \max_{\pi \in \mathrm{PD}(A)} \min_{o \in O} \sum_{a \in A} R_{o,a} \pi_a$$

Markov decision process (MDP)

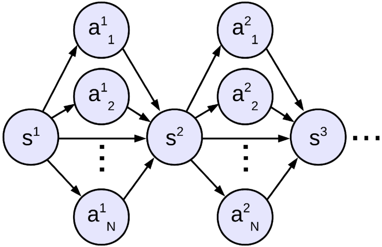
Value function

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} \cdot R_t \mid S_0 = s\right]$$

Bellman optimality operator T

$$(TQ)(s,a) = r(s,a) + \gamma \cdot \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q\left(S', a'\right) \mid S' \sim P(\cdot \mid s,a)\right]$$

Markov Games

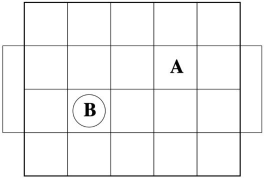


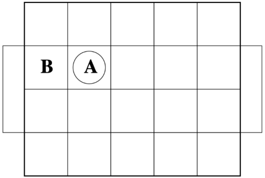
- Define the Bellman operators T (for two-player zero-sum game) by

$$(TQ)(s,a,b) = r(s,a,b) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,a,b)} \left\{ \max_{\pi' \in \mathcal{P}(\mathcal{A})} \min_{\nu' \in \mathcal{P}(\mathcal{B})} \mathbb{E}_{a' \sim \pi', b' \sim \nu'} \left[Q\left(s', a', b'\right) \right] \right\}.$$

- When action $|A| = 1$, MGs is reduced to MDP.
  When state $|S| = 1$, MGs is reduced to matrix games.

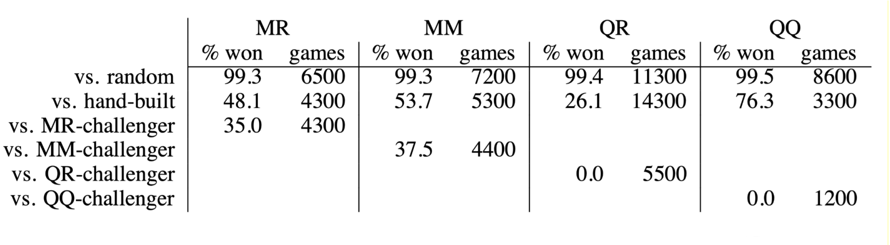
The Game of Soccer





- The two players choose one of 5 actions on each turn: N, S, E, W, and stand.
- The agent must use a probabilistic policy to breaking an unknown defender.

Results of the Game



	MR		MM		QR		QQ	
	% won	games	% won	games	% won	games	% won	games
vs. random	99.3	6500	99.3	7200	99.4	11300	99.5	8600
vs. hand-built	48.1	4300	53.7	5300	26.1	14300	76.3	3300
vs. MR-challenger	35.0	4300						
vs. MM-challenger			37.5	4400				
vs. QR-challenger					0.0	5500		
vs. QQ-challenger							0.0	1200

- Minimax-Q (M), Q-learning (Q) and random (R).
- Surprisingly, the QQ policy did so well against the hand-built opponents.
- Trained by Q-learning did significantly worse (due to deterministic policies).
- Minimax criterion allows the agent to converge to a “safe” strategy.

1 From Matrix Games and MDP to Markov Games

2 RL in Cooperative Multiagent Systems

The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems
-*MAAAI/IAAI* (Caroline Claus and Craig Boutilier, 1998)

Reinforcement Learning of Coordination in Cooperative Multi-agent Systems
-*Adaptive Agents and Multi-Agent Systems II*
(Spiros Kapetanakis and Daniel Kudenko, 2004)

Multiagent Reinforcement Learning (MARL) Algorithm I

N-player cooperative repeated games

	$a0$	$a1$
$b0$	x	0
$b1$	0	y

	$a0$	$a1$	$a2$
$b0$	10	0	k
$b1$	0	2	0
$b2$	k	0	10

	$a0$	$a1$	$a2$
$b0$	11	-30	0
$b1$	-30	7	6
$b2$	0	0	5

A strategy profile Π is a Nash equilibrium iff $\Pi(i)$ is a best response to Π_{-i}. An equilibrium (or joint action) is optimal if no other has greater value.

Learning in coordination Games

For each agent j, i assumes j plays action $a^j \in A_j$ with probability $Pr^j_{a^j} = \frac{C^j_{a^j}}{\sum_{b^j \in A_j} C^j_{b^j}}$.

This simple adaptive strategy will converge to an equilibrium.

Multiagent Reinforcement Learning (MARL) Algorithm II

Q-value (for convergence)

Updates estimate $Q(a)$ as $Q(a) \leftarrow Q(a) + \lambda(r - Q(a))$ (Generally, $Q(a, s)$ is taken).
If λ is decreased "slowly" and a is sampled infinitely, Q-learning will converge.

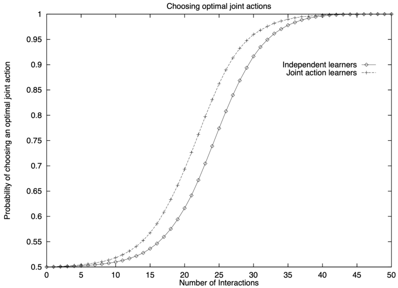
Exploration/exploitation problem (for optimality)

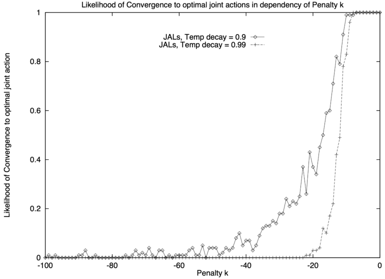
Exploitive exploration: choose its best estimated action with probability p_x.
Boltzmann exploration: action a is chosen with probability $e^{Q(a)/T} / \sum_{a'} e^{Q\left(a'\right)/T}$.

Learning ways

Independent learner (IL, with a^i) and **joint action learner** (JAL, with a).
For JAL, agent i assesses $EV\left(a^i\right) = \sum_{a^{-i} \in A_{-i}} Q\left(a^{-i} \cup \left\{a^i\right\}\right) \prod_{j \neq i} \left\{\Pr^i_{a^{-i}[j]}\right\}$.

Independent Learner vs Joint Action Learner





- Let $x = y = 10$ and $T = 16 * 0.9^t$. Both ILs and JALs use Boltzmann exploration.
- JALs do perform better, while convergence is not enhanced dramatically.
- If $k = -100$, initial exploration will find the first and third to be unattractive.

Conditions for Multiagent Q-learning Convergence

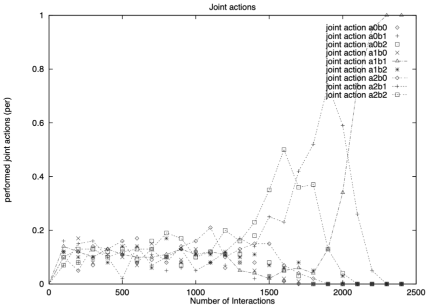


Figure: Results for climbing game.

Conditions for convergence

E_t denotes the probability of a equilibrium. For both ILs and JALs, for all $t > T(\delta, \epsilon)$, $\Pr\left(|E_t - 1| < \varepsilon\right) > 1 - \delta$, when

1. $\sum_{t=0}^{T} \lambda_t = \infty$ and $\sum_{t=0}^{T} \lambda_t^2 < \infty$.
2. Samples each actions infinitely often.
3. $P_t^i(a) \neq 0$.
4. $\lim_{t \to \infty} P_t^i(X_t) = 0$ for estimated nonoptimal action X_t.

Biasing Exploration Strategies for Optimality

Some myopic heuristics algorithms

- **Optimistic Boltzmann (OB)**
  $\max_{\Pi_{-i}} Q(\Pi_{-i}, a_i)$ as the value of a_i.
- **Weighted OB (WOB)**
  using $\max_{\Pi_{-i}} Q(\Pi_{-i}, a_i) \cdot Pr_i$.
- **Combined**
  using $\rho \max_{\Pi_{-i}} Q(\Pi_{-i}, a_i) + (1-\rho) EV(a_i)$.

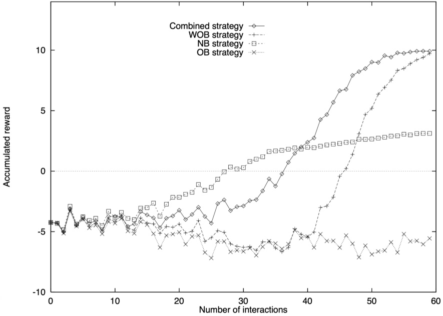
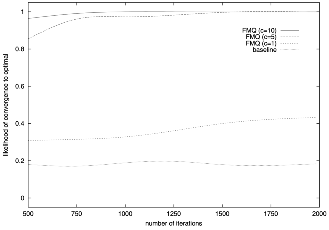
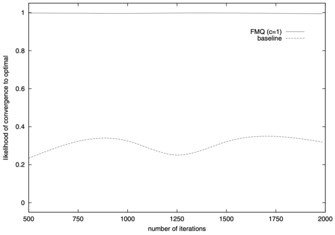
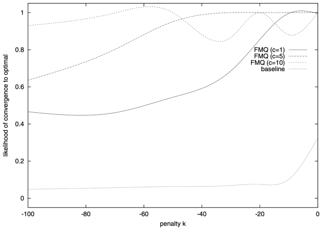


Figure: Results for penalty game $k = -10$.

More Results about ILs





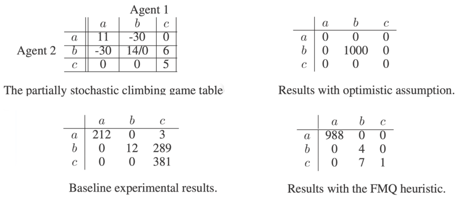


Frequency Maximum Q Value (FMQ) heuristic (Kapetanakis et al., 2004)

$$\mathrm{EV}(\alpha) = Q(\alpha) + c * \mathsf{freq}\,(\max \mathrm{R}(\alpha)) * \max \mathrm{R}(\alpha).$$

- Let $T(x) = e^{-sx} T_0 + 1$. The results about climbing game and penalty game.

Compare FMQ to Optimistic Assumption



		Agent 1		
		a	*b*	*c*
Agent 2	*a*	11	-30	0
	b	-30	14/0	6
	c	0	0	5

The partially stochastic climbing game table

	a	*b*	*c*
a	0	0	0
b	0	1000	0
c	0	0	0

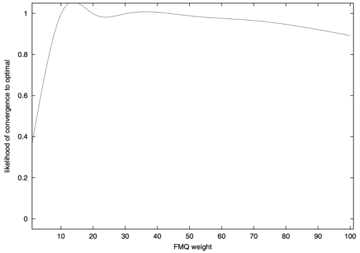
Results with optimistic assumption.

	a	*b*	*c*
a	212	0	3
b	0	12	289
c	0	0	381

Baseline experimental results.

	a	*b*	*c*
a	988	0	0
b	0	4	0
c	0	7	1

Results with the FMQ heuristic.



- *Optimistic assumption* only succeeds in solving the deterministic climbing game.
- By setting the FMQ weight too high, the probabilities for action selection are influenced too much towards the action with the highest FMQ value.
- For a fully stochastic climbing game, both heuristics perform poorly.

Thanks!