The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems

2/1

強化学習勉強会

大録　誠広

Abstract & 1. Introduction

- 複数のプレイヤーがいるような状況でどのように協調が発生するか⇒強化学習(RL)でモデル化
- 他のプレイヤーの存在に気付いていないケース⇔joint actionのvalueと相手の戦略を明示的に学ぼうとするケース
- ゲームの構造と探索戦略がNash均衡への収束にどのように影響するかを調べた
- 最適な均衡への収束の確からしさを増大させる”optimistic”な探索戦略を提案した

2. Preliminary Concepts and Notation
2.1 Single Stage Games

- Nプレイヤーの繰り返し調整ゲーム
- distributed bandit problemとして扱う

モデルの定式化

$$\text{a collection of n players} : \alpha$$

$$\text{each agent} : i \in \alpha$$

$$\text{a finite set of individual actions} : A_i$$

$$\text{the set of joint actions} : \mathcal{A} = \times_{i \in \alpha} A_i$$

$$\text{each joint action} : a \in \mathcal{A}$$

$$\boxed{\text{expected reward} : R(a)}$$

cooperative since each agent's reward is drawn from the same distribution

戦略と均衡の定義

$$\text{randomized stragey for agent } i : \pi \in \Delta(A_i)$$

$$\text{probability of agent } i \text{ selecting action } a^i \in A_i : \pi(a^i)$$

$$\text{a strategy profile} : \Pi = \{\pi_i : i \in \alpha\}$$

Given a profile Π_{-i}, a strategy π_i is a *best response* for agent i
if the expected value of the strategy profile $\Pi_{-i} \cup \{\pi_i\}$
is maximal for agent i

the strategy profile Π is a *Nash equilibrium* iff $\mathit{\Pi}[i]$(*i's component of* $\mathit{\Pi}$)
is a best response to Π_{-i}, for every agent i

例

ナッシュ均衡: 自分から逸脱する誘因が無い
Optimal : 二人にとって最も利得が高い

	a0	a1
b0	x	0
b1	0	y

$x > y > 0$

Nash equilibria

Optimal

2. Preliminary Concepts and Notation
2.2 Learning in Coordination Games

- 最適な均衡が複数ある場合⇒行動選択は難しくなる
- 協調行動は同じプレイヤー間の

繰り返しゲームの結果学習され得る

- 収束を保証する学習モデル：仮想プレイ

相手が過去に取った行動の割合と同じ確率で

次の行動を選ぶという信念

	a0	a1
b0	x	0
b1	0	y

$x = y > 0$

Nash equilibria

Optimal

2. Preliminary Concepts and Notation
2.3 Reinforcement Learning

- Joint actionに紐付く利得の情報を知らない場合⇒強化学習を用いて過去の経験から類推することが可能
- Stateless のQ学習（Q学習というより、基本的なstochaostic approximation technique）

action : a

reward : r

Q value : $Q(a)$

An agent updates its estimate $Q(a)$ based on sample $< a, r >$ as follows :

$$Q(a) \leftarrow Q(a) + \lambda(r - Q(a))$$

Exploitation vs Exploration

- Nonoptimal な行動を取る確率 ⇒Boltzmann exploration など

action a is chosen with prob. $\dfrac{e^{Q(a)/T}}{\sum_{a'} e^{Q(a')/T}}$

2.3 Reinforcement Learning（続き）

- 一般にmultiagentの(nonstationallyな環境での)Q学習は難しい。Q-valueの収束は保証されていない

Q学習をMultiagentに適用可能な手法 : MARL (IL)アルゴリズムとJAL

2.3 Reinforcement Learning（続き）

- MARL もしくはIndependent Learner(IL)
- プレイヤーは自分の行動とrewardを経験$< a^i, r >$としてQ学習

2.3 Reinforcement Learning（続き）

- Joint Action Learner (JAL)
- 自分と相手の行動を経験$< a , r >$として学習
- 相手は現在の自分のbeliefに沿って行動すると考える

$$EV(a^i) = \sum_{a^{-i} \in A_{-i}} Q(a^{-i} \cup \{a^i\}) \prod_{j \neq i} \{\mathrm{Pr}^i_{a^{-i}[j]}\}$$

2.3 Reinforcement Learning（続き）

AgentのExperience

- Independent Learner (IL) の場合: $< a^i, r >$
- Joint Action Learner(JAL)の場合: $< a\ , r >$

どちらもPartially observable model $< a^i, o, r >$ の特殊なケース

Aのaction setが$\{a_0, a_1\}$ Bのaction setが$\{b_0, b_1\}$である時, AのQ値は

- Independent Learner (IL) の場合:

$Q(a_0), Q(a_1)$の2通り

- Joint Action Learner(JAL)の場合:

$Q(< a_0, b_0 >), Q(< a_0, b_1 >), Q(< a_1, b_0 >), Q(< a_1, b_1 >)$の4通り

3. Comparing Independent and Joint-Action Learners

	a0	a1
b0	10	0
b1	0	10

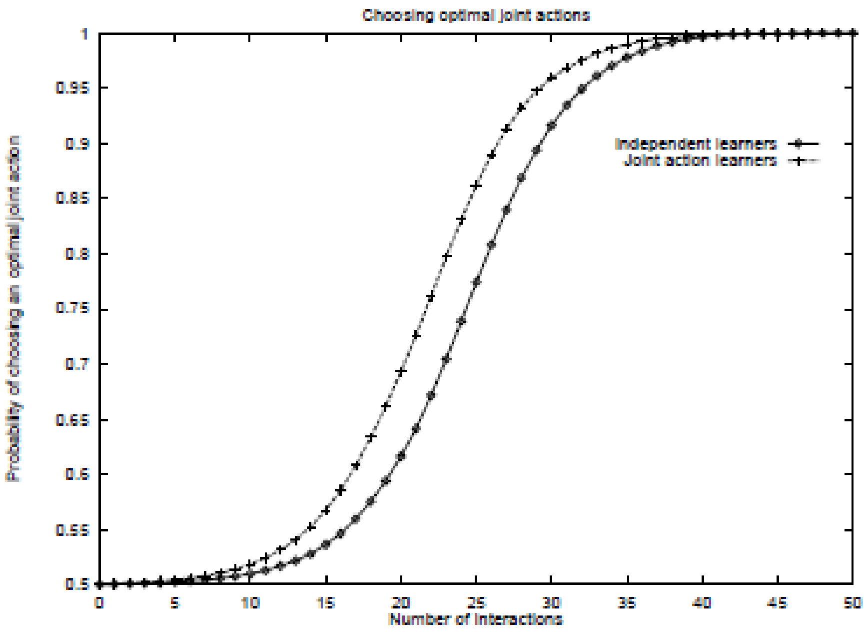

Figure 1: Convergence of coordination for ILs and JALs (averaged over 100 trials).

- IL, JALで比較（どちらもBoltzmann exploration ）
- JALの方がわずかに早く収束
- お互いにILをやっているのと同じようなものである&exploration strategyにより

4. Convergence and Game Structure

- ゲームの構造がより複雑な場合
- どこに収束する？

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

Penalty Game
$k \leq 0$

4. Convergence and Game Structure

- ゲームの構造がより複雑な場合
- どこに収束する？

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

Penalty Game
$k \leq 0$

Nash equilibria

4. Convergence and Game Structure

- ゲームの構造がより複雑な場合
- どこに収束する？

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

Penalty Game
$k \leq 0$

Nash equilibria

Optimal

4. Convergence and Game Structure

- k = -100だったら？

	a0	a1	a2
b0	10	0	-100
b1	0	2	0
b2	-100	0	10

Optimalではない
均衡が選択される

4. Convergence and Game Structure

- kに対する応答

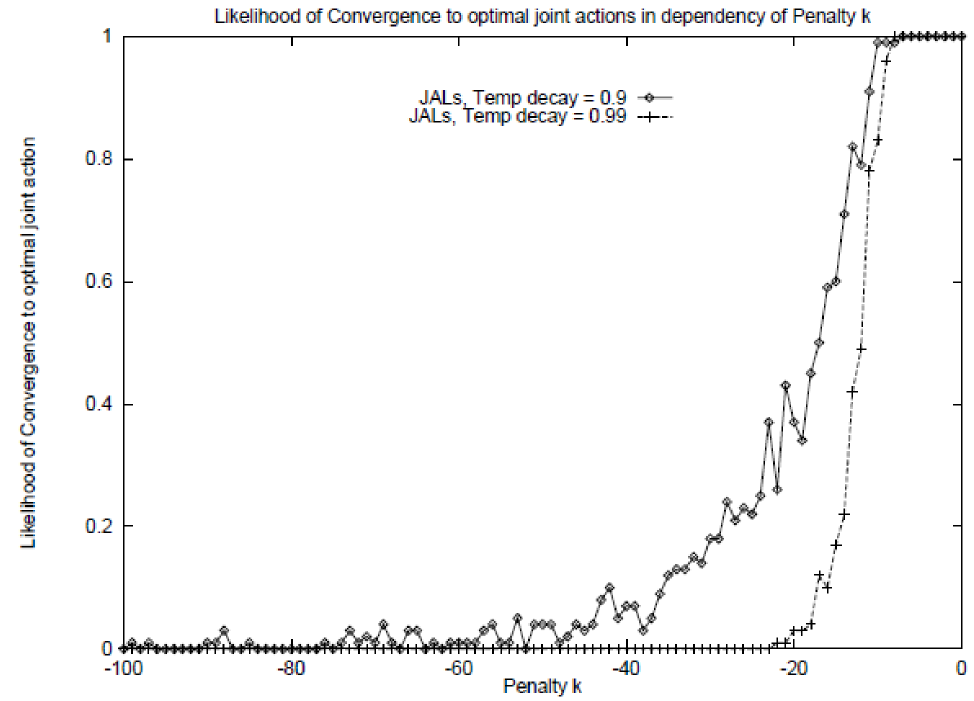

Figure 2: Likelihood of convergence to opt. equilibrium as a function of penalty k (averaged over 100 trials).

4. Convergence and Game Structure(続き)

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

Climbing Game

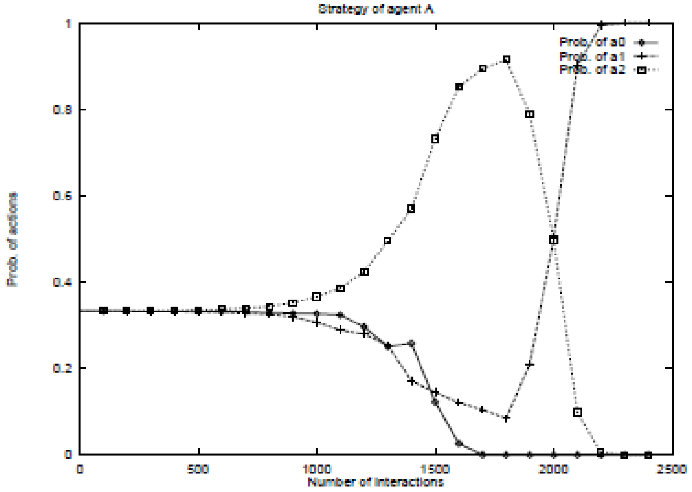

Figure 3: A's strategy in climbing game

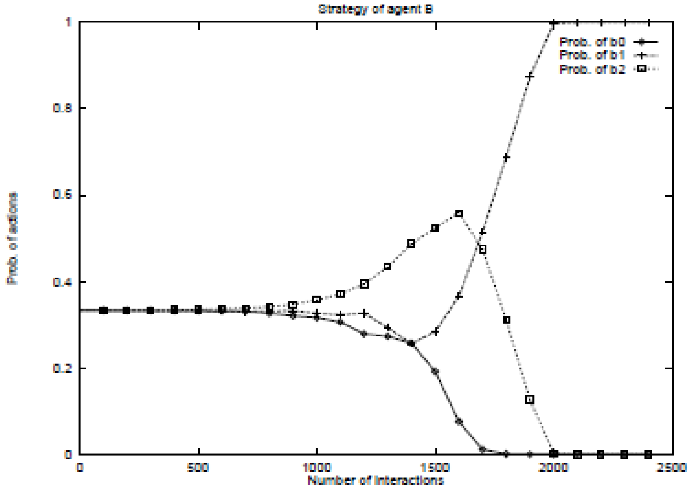

Figure 4: B's strategy in climbing game

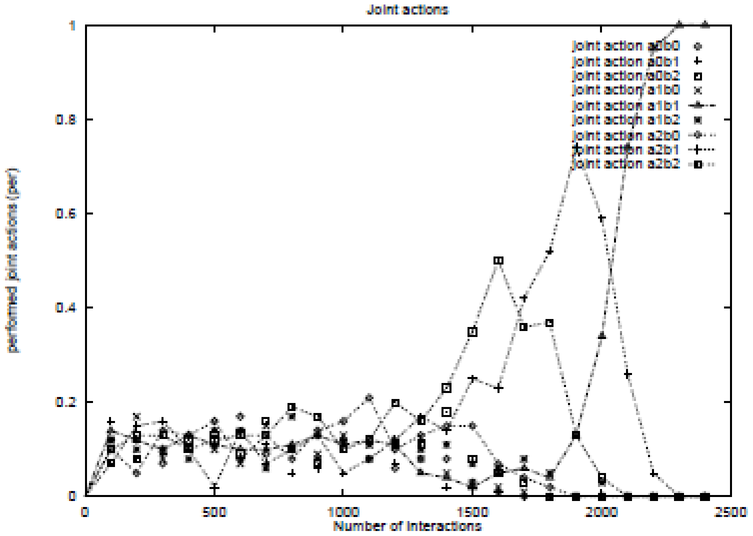

Figure 5: Joint actions in climbing game

4. Convergence and Game Structure(続き)

- 収束の要件

- The learning rate λ decreases over time such that $\sum_{\lambda=0}^{t} \lambda = \infty$ and $\sum_{\lambda=0}^{t} \lambda^2 < \infty$.
- Each agent samples each of its actions infinitely often.
- The probability $P_t^i(a)$ of agent i choosing action a is nonzero.
- Each agent's exploration strategy is exploitive. That is, $\lim_{t\to\infty} P_t^i(X_t) = 0$, where X_t is a random variable denoting the event that some nonoptimal action was taken based on i's estimated values at time t.

Theorem 1 *Let E_t be a random variable denoting the probability of a (deterministic) equilibrium strategy profile being played at time t. Then for both ILs and JALs, for any $\delta, \varepsilon > 0$, there is an $T(\delta, \varepsilon)$ such that*

$$\Pr(|E_t - 1| < \varepsilon) > 1 - \delta$$

for all $t > T(\delta, \varepsilon)$.

5. Biasing Exploration Strategies for Optimality

MARL(IL)は、最適な均衡への収束は保証しない
JALなら、より”optimistic”な探索戦略（myopic heuristics）を取ることによってoptimalな均衡へ収束するlikelihoodを高めることが出来る

Optimistic Boltzmann (OB): For agent i, action $a_i \in A_i$, let $MaxQ(a_i) = \max_{\Pi_{-i}} Q(\Pi_{-i}, a_i)$. Choose actions with Boltzmann exploration (another exploitive strategy would suffice) using $MaxQ(a_i)$ as the value of a_i.

Weighted OB (WOB): Explore using Boltzmann using factors $MaxQ(a_i) \cdot \Pr_i(\text{optimal match } \Pi_{-i} \text{ for } a_i)$.

Combined: Let $C(a_i) = \rho\ MaxQ(a_i) + (1-\rho)EV(a_i)$, for some $0 \le \rho \le 1$. Choose actions using Boltzmann exploration with $C(a_i)$ as value of a_i.

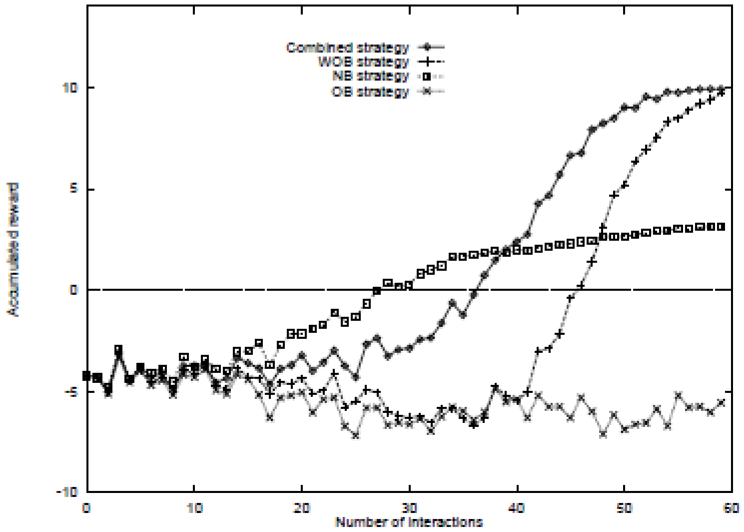

Figure 6: Sliding avg. reward in the penalty game

6. Concluding Remarks

- 複数のプレイヤーがいる場合、Q学習（による最適方策への収束）はプレイヤーが一人の時に比べて頑健ではない
- 複雑なゲーム的状況においては経験則による新しい探索手法が有効である
- 今後の方向性
  - Q学習が適用される、複数の状態を持つ逐次的な問題に対する応用
  - 状態と行動の集合がより大きい時（特に、プレイヤーの数に対して状態の数が指数関数的に増加する場合）の一般化
  - 仮想プレイでの収束が知られている他の状況（ゼロサムゲームなど）での応用