論　文

Technical Papers

マルチエージェント強化学習の方法論
—Q-Learningと Profit Sharingによる接近—

Methodology in Multi-Agent Reinforcement Learning
—Approaches by Q-Learning and Profit Sharing—

荒井　幸代*　宮崎　和光*　小林　重信*
Sachiyo Arai　Kazuteru Miyazaki　Shigenobu Kobayashi

* 東京工業大学大学院総合理工学研究科
Graduate School of Interdisciplinary Science and Engineering, Tokyo Institute of Technology, Yokohama 226-8502, Japan.

1997年7月25日　受理

Keywords: multi-agent systems, reinforcement learning, pursuit problem, Q-learning, profit sharing.

Summary

Most of multi-agent systems have been developed in the field of Distributed Artificial Intelligence (DAI) whose schemes are based on plenty of pre-knowledge of the agents' world or organized relationships among the agents. However, these kind of knowledge would not be always available. On the other hand, multi-agent reinforcement learning is worth considering to realize the cooperative behavior among the agents with little pre-knowledge.

There are two main problems to be considered in multi-agent reinforcement learning. One is the *uncertainty of state transition* problem which is owing to concurrent learning of the agents. And the other is the *perceptual aliasing* problem which is generally held in such a world. Therefore, the robustness and flexibility are essential for the multi-agent reinforcement learning toward these two problems.

In this paper, we evaluate Q-learning and Profit Sharing as the method for multi-agent reinforcement learning through some experiments. We take up the *Pursuit Problem* as one of the multi-agent world. In the experiments, we do not assume the existence of any pre-defined relationship among agents or any control knowledge for cooperation. Learning agents do not share sensation, episodes and policies. Each agent learns through its own episodes independent of the others.

The result of experiments shows that cooperative behaviors emerge clearly among Profit Sharing hunters who are not influenced by concurrent learning even when the prey has the certain escaping way against the hunters. Moreover, they behave rational under the perceptual aliasing areas. On the other hand, Q-learning hunters can not make any policy in such a world. Through these experiments, we conclude that Profit Sharing has the good properties for multi-agent reinforcement learning because of its robustness for the change of other agents' policies and the limitation of agent's sensing abilities.

1. はじめに

マルチエージェント系における協調的行動の実現は，工学的及び認知科学的観点からたいへん興味ある話題であり，これまでにさまざまな研究がなされてきた[Werner 91]．マルチエージェント系への接近法は，トップダウンアプローチとボトムアップアプローチの二つに大別される．前者に基づく研究は，DAI(Distributed Artificial Intelligence)のコミュニティを中心に80年代初めより今日まで活発に展開されている[Smith 81, Weiss 95]．DAIでは，各エージェントの行動戦略及びエージェント間の通信規約を事前に知識としてシステムに組み込むことを前提としているが，これらの知識は問題領域に依存することから，エキスパートシステムと同様に，知識獲得という重大な問題に直面せざる

を得ない [Sen 95].

一方，後者に基づく研究は，強化学習の発展を契機として，90 年代に入って始まった新しいパラダイムであり，エージェントの行動戦略を学習によって獲得させようとする点において DAI 的接近とは大きく異なる．

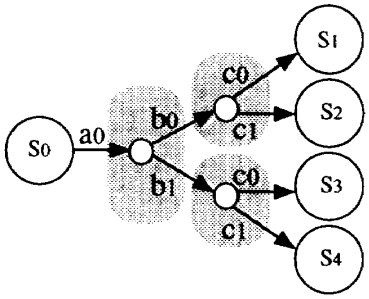
マルチエージェント強化学習は，[Whitehead 91] を嚆矢として，[Ono 96, Sen 95, Tan 93, Weiss 93] などの先駆的研究がある．具体的方法として，[Ono 96, Tan 93, Whitehead 91] は Q-learning を，[Sen 95, Weiss 93] は Classifier System を，それぞれベースにしている．

Q-learning は，マルコフ決定過程 (Markov Decision Processes：MDPs) の下で，最適解に収束することが [Watkins 92] によって証明されているが，マルチエージェント系の環境は，一般に，MDPs ではないことから，Q-learning に基づく接近は，実際には，いくつかの本質的問題に直面することが予想される．

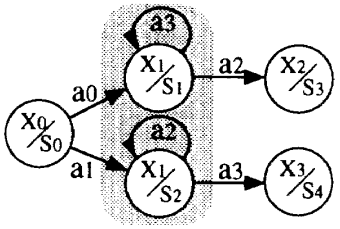
Classifier System における強化学習アルゴリズムとして，Bucket Brigade[Holland 86] や Profit Sharing[Grefenstette 88] がよく使われている．Bucket Brigade と Profit Sharing は，いずれも市場経済のアナロジーから生まれたヒューリスティックな方法であるが，Profit Sharing については，[宮崎 94] の合理性定理 (無効ルールの抑制及び報酬プランの獲得) によって安定的な挙動が保証されている．これらの定理は MDPs の前提を必要とせず，非 MDP クラスに類別されるマルチエージェント系に対しても自然に適用できることに留意されたい．

マルチエージェント強化学習の方法論として，どの手法が妥当かつ適切であるかについて，まだ客観的な判断を下すのに必要な知見やデータが蓄積されていないのが現状である．[Sen 95] は，ロボットナビゲーション問題及び資源共有問題を使って，マルチエージェント強化学習の有効性を確認する実験を行い，Q-learning, Bucket Brigade, Profit Sharing のいずれも，マルチエージェント系に対して適用可能であると結論づけている．しかし，[Sen 95] が取り上げた二つの問題は，強化学習に固有の特徴である環境に対する入力の不完全性や報酬の遅れを同時に扱うものではなく，この結論は少し割り引いて受けとめる必要がある．

本論文の目的は，Q-learning と Profit Sharing の二つの強化学習を取り上げ，マルチエージェント系の代表的なベンチマークである追跡問題 [Gasser 89] への適用を通じて，マルチエージェント強化学習の方法論としてのこれらの手法の特徴と限界を明らかにすることにある．



(a) 他エージェントの存在による状態遷移の不確定性



(b) 感覚入力の制限による不完全知覚

ai：エージェントAの行動
bi：エージェントBの行動
ci：エージェントCの行動
Sj：環境の状態
Xk：エージェントのk番目の感覚入力

図 1 マルチエージェント系における課題

以下，**2 章**では，マルチエージェント強化学習における問題の所在を指摘した上で，Q-learning による接近及び Profit Sharing による接近について考慮すべき事項を議論する．**3 章**では，一般的な追跡問題を定義した後，この問題を 5 つのケースに分類し，各ケースの特徴を強化学習の視点から論じる．**4 章**では，Q-learning と Profit Sharing の比較実験について，実験の方法，実験の結果及び考察を述べる．**5 章**は結論で，本研究の成果をとりまとめ，今後の研究課題について論じる．

2. 強化学習によるマルチエージェント系への接近

2・1 マルチエージェント強化学習における問題の所在

本論文では，各エージェントが独立に学習するマルチエージェント強化学習を対象とする．この設定の下では，複数エージェントの同時学習 (concurrent learning[Sen 95]) によって生じる状態遷移の不確定性，及び各エージェントの感覚入力の限界によって生じる不完全知覚の二つの問題が重要である．

i. 状態遷移の不確定性

図 1(a) の状態 S_0においてエージェント A がある行動 a_0 を取った後，続くエージェント Bと Cの行動

如何によって状態遷移先はS_1～S_4のいずれかとなる．ここで，各エージェントが学習途中であれば，S_0から$\{S_1, S_2, S_3, S_4\}$への状態遷移確率は動的に変化するし，さらには状態遷移先も変化する場合が起こり得る．これを状態遷移の不確定性 (uncertainty of state transition) 問題と呼ぶ．複数のエージェントがそれぞれ独立に学習する場合，状態遷移が不確定となり，環境は非MDPのクラスになることに注意しなければならない．

ii. 不完全知覚

エージェントの感覚入力が制限されているとき，図1(b)に示すように，エージェントAは異なる状態S_1とS_2において同じ感覚入力X_1を受け取るために，両者を区別することはできない．ここで，状態S_1においてa_2が好ましい行動で，a_3は好ましくない行動，状態S_2においてa_3が好ましい行動で，a_2は好ましくない行動だとすると，エージェントAは感覚入力X_1の下ではつねに好ましい行動を出力することは不可能になる．これは不完全知覚 (perceptual aliasing) 問題と呼ばれる [Whitehead 90]．

不完全知覚に対処する第1の方法は，過去の行動系列を利用して不完全知覚状態を識別することであるが，各エージェントが独立に学習するマルチエージェント系では，相互に過去の行動系列を参照するという設定は不自然であり，この方法を採用することはできない．第2の方法は，確率的政策によって不完全知覚状態から確率的に脱出することであるが，確率的政策の利用可能性は強化学習アルゴリズムに依存することに注意しなければならない．

マルチエージェント系に強化学習を適用する際，上記の二つの問題に対して，どこまで対応できるかを定性的かつ定量的に把握しておくことが必要であり，本論文の動機はここにある．

本論文で比較対象として取り上げるQ-learningとProfit Sharingは，それぞれ環境同定型と経験強化型の代表的手法である [山村 95]．これらは共に状態遷移の不確定性及び不完全知覚を伴う問題に対してどのような挙動を示すのかについては明らかにされていない．そこで，本論文ではこれらの挙動を明らかにすることを考える．

2・2 Q-learningによる接近

Q-learningは，報酬に至るエピソードの各ステップごとに，以下の(1)および(2)式を用いて，状態と行動の各組に対するQ値を更新することを繰り返して政策を形成する．

$$Q(x,a) \leftarrow Q(x,a) + \alpha(r + \gamma V(y) - Q(x,a)) \tag{1}$$

$$V(x) = max_{b \in actions} Q(x,b) \tag{2}$$

ここで，xは現在の状態，aは選択行動，yは遷移後の状態，rは報酬値，γ（$0 \leq \gamma < 1$）は割引率，α（$0 < \alpha < 1$）は学習定数である．

[Watkins 92]により，環境がMDPsであるとき，学習定数αをある条件が満たされるように調整すれば，Q-learningは最適政策に収束することが証明されている．しかし，マルチエージェント系の環境は，一般に，非MDPとなるので，Q-learningの収束性は保証されないが，上述の状態遷移の不確定性および不完全知覚の影響が大でない限り，局所的政策への収束は期待できる．なお，その場合でも，学習係数αを十分小さく設定することが必要である．

Q-learningの学習過程における行動選択の方法として，Boltzmann選択，max選択とランダム選択の組み合わせ，ルーレット選択などがあるが，Boltzmann選択は，一般に広く用いられており [Ono 96, Tan 93, Whitehead 90]，著者らの経験でも追跡問題に限らず，非決定性を含む問題に対して良い性能を示すことが確認されている．

さらに，学習後の行動選択の方法として，MDPs環境ではmax選択により最適政策が得られるが，不完全知覚などに起因する非MDP環境では決定的政策より確率的政策のほうが有効との報告 [Singh 94] もあり，max選択よりはBoltzmann選択のほうが有効と考えられる．

以上の理由及び予備的実験の結果を踏まえて，本論文ではQ-learningの学習過程及び学習後の行動選択として(3)式で示されるBoltzmann選択を採用する．

$$p(a_i|x) = \frac{e^{Q(x,a_i)/T}}{\sum_{k \in actions} e^{Q(x,a_i)/T}} \tag{3}$$

2・3 Profit Sharingによる接近

Profit Sharingは，報酬に至るエピソードにおける感覚入力xと行動aの対からなるルール系列を記憶しておき，報酬が得られた時点で系列上のルールを次式に従って強化する．

$$w(x_i,a_i) \leftarrow w(x_i,a_i) + f(r,i) \tag{4}$$

ここで，$w(x_i,a_i)$はエピソード系列上のi番目のルールの重み，rは報酬値，fは強化関数である．

[宮崎 94]により，強化関数について政策の局所的合

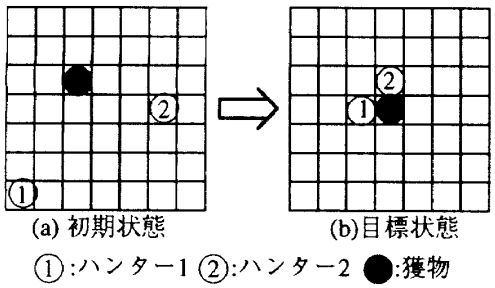


図2 追跡問題

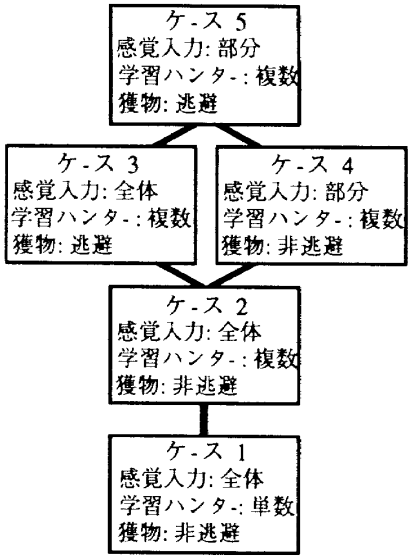


図3 強化学習からみた追跡問題の階層

理性を保証する必要十分条件が証明されている．[宮崎 94] の合理性定理は，最適性を保証するものではないが，MDPs の仮定を必要としないので，マルチエージェント系のような非 MDP 環境に対しても適用できる点に特徴がある．

Profit Sharing の学習過程における行動選択の方法としては，ルーレット選択が良い性能を示すことが経験的に知られている．また，ルーレット選択は確率的政策を自然に実現する枠組みでもあり，非 MDP 環境における学習後の行動選択としても有効である．

以上の理由及び予備的実験の結果を踏まえて，本論文では Profit Sharing の学習過程及び学習後の行動選択としてルーレット選択を採用する．

3. 問 題 設 定

3・1 追 跡 問 題

図2に示す $n \times n$ 格子状トーラスの環境を設定し，ここに複数のハンターエージェントと獲物エージェントをランダムに配置したものを初期状態とする．各エージェントは，予め決められた順番で行動し，上下左右の方向に1コマ進むかまたは停止の行動を一つ選択する．複数のエージェントが同一場所に存在することは許さない．

ハンターの視界は $m \times m (n \geq m)$ で，自分の周囲 $m^2 - 1$ コマが見える．以下の説明では，$n = m$ のときを感覚入力が全体的と呼び，$n > m$ のときを部分的と呼ぶ．図2(b) のように，すべてのハンターが獲物に隣接したときを目標状態とし，すべてのハンターに報酬を与える．

獲物の行動パターンとして，非逃避的行動または逃避的行動のどちらかを設定する．非逃避的行動では，獲物はハンターの動きとは無関係にランダムに行動を選択する．逃避的行動では，獲物の視界は 5×5 （ただし，斜め方向は死角）とし，視界内にあるハンターから遠ざかる方向に逃げる行動を選択する．視界に何も存在しない場合にはランダムに行動を選択する．獲物は学習しない．

各ハンターは，通信や情報の共有なしに，報酬を唯一の手がかりとして，それぞれ独立に強化学習の枠組みに従って政策を学習する．獲物を捕獲するための協調的行動は学習の結果として出現することに注意されたい．

3・2 追跡問題の分類

2・1 節で議論した問題が強化学習アルゴリズムの収束性などにどのような影響を与えるのかを調べるために，追跡問題を図3に示す階層に分類して，**4 章**での実験を計画する．すなわち，

- ハンターの感覚入力：全体的または部分的，
- 学習するハンター：単数または複数，
- 獲物の行動パターン：非逃避的または逃避的，

の組み合わせによって，以下の5つのケースの問題を設定する．

ケース1 〔**感覚入力／全体的，学習ハンター／単数，獲物／非逃避的**〕：ケース1はシングルエージェントの強化学習問題であり，不完全知覚が生じないことから，学習ハンターにとって環境は MDPs である．

ケース2 〔**感覚入力／全体的，学習ハンター／複数，獲物／非逃避的**〕：学習ハンターが複数になっている点がケース1と異なる．この場合，複数のハンターが同時かつ独立に学習を行うので，学習過程において状態遷移の不確定性が生じ，学習ハンターにとって環境は非 MDP となる．ただし，他の学習ハンターが何らかの政策に収束すれば環境は MDPs となる．

ケース3｛**感覚入力／全体的，学習ハンター／複数，獲物／逃避的**｝：獲物の逃避的行動は，一般に，ハンターの学習を妨げる要因として働くことから，ケース2に比べて学習過程における状態遷移の不確定性の度合がより大きくなることが予想される．

ケース4｛**感覚入力／部分的，学習ハンター／複数，獲物／非逃避的**｝：ケース2の状態遷移の不確定性に加えて，不完全知覚の問題が新たに生じるので，ケース2より難しい問題となるが，獲物は非逃避的であることから，ケース3に比べて必ずしも難しいとはいえない．なお，[Ono 96, Tan 93] はこのケースを扱っている．

ケース5｛**感覚入力／部分的，学習ハンター／複数，獲物／逃避的**｝：学習を妨げる要因をすべて含んでおり，追跡問題としてはもっとも難しいケースである．ケース4に比べて状態遷移の不確定性の度合はより大きくなり，それに伴い不完全知覚が生じる度合も大きくなると予想される．

以上を要約すると，状態遷移の不確定性には複数の学習ハンターの存在と獲物の逃避行動が直接関係する．一方，不完全知覚には視覚の制限による状態観測（感覚入力）の部分性が直接関係する．さらに，不完全知覚の度合は状態遷移の不確定性の度合の影響を強く受ける可能性があることに注意されたい．

4. 実　　験

本章では，**3章**で設定したケース1～5のそれぞれに対して，学習ハンターの学習アルゴリズムとしてQ-learningまたはProfit Sharingを実装し，獲物を捕獲するための協調的行動を学習するプロセスを実験により調べ，結果について考察を加える．

4・1 実 験 方 法

7×7格子状トーラスの環境にハンター2人と獲物1匹をランダムに配置する．学習するハンターの視界は，ケース1～3では7×7，ケース4と5では5×5とする．ケース1のみハンターの1人は学習しない．このハンターは他ハンターの位置や動きとは無関係に獲物に近づく行動を選択する．そのような行動が複数ある場合は，上，右，下，左の順に行動を選択する．

エピソード（初期状態から始まり，目標状態に到達するまで）ごとに，初期状態でのハンターと獲物はランダムに配置する．10万エピソードの繰り返しを1試行とする．

Q-learningでは，Q値の初期値を0.1，報酬を1，割引率γを0.9，学習定数αを0.04，Boltzmann選択における温度Tを0.2に，それぞれ設定した．Profit Sharingでは，各ルールの初期重みを0.1，報酬を1に設定し，公比0.2の等比減少関数を強化関数に採用した．これらの値は，経験的な知見を参考に，予備的実験によって最も高い性能を示すことを確認の上，設定した．

4・2 実験結果および考察

各ケースごとに，Q-learningとProfit Sharingによって，それぞれ10試行実験を行った．それぞれの学習曲線および学習後のエピソード長（報酬を獲得するまでのステップ数）の平均と分散を図4に示す．以下，実験結果について考察する．

〔1〕 **感覚入力が全体の場合（ケース1～3）**

i. ケース1の結果

学習後のQ-learningとProfit Sharingのエピソード長は同程度であり，共にほぼ最適な政策を獲得したものと判断される．このケースは学習ハンターにとってMDPs環境であり，Q-learningにより最適政策が獲得されるのは当然であるが，局所的合理性の裏付けしかないProfit SharingがQ-learningと同等の性能を示していることは注目される．

学習の立ち上がりについては，予想されたとおり，Profit SharingのほうがQ-learningに比べて素早いが，これは報酬の直接フィードバックによる経験強化というProfit Sharingの特徴を反映したものと考えられる．以下のすべてのケースにおいて同様の傾向が観察される．

ii. ケース2の結果

学習ハンターが2人であるために，**3・2節**で論じたように，学習過程において状態遷移の不確定性が生じ，それが学習の収束速度を遅らせる原因となっていると推測される．特に，Q-learningの学習初期においてその傾向が著しい．Profit Sharingでは，(4) 式に従い，各ルールの重みは加算されるだけなので，状態遷移の不確定性に対して頑健で，不確定性を吸収する特徴をもつ．一方，Q-learningでは，(1)，(2) 式に従ってQ値が更新されることから，特に，(2) 式のmax演算においてQ値の変動に対して敏感に反応してしまうので，状態遷移の不確定性に対し頑健ではないといえる．

学習後の性能については，Q-learningもProfit Sharingも同程度であり，ケース1との比較から，共にほぼ最適な政策を獲得することに成功していると判断される．

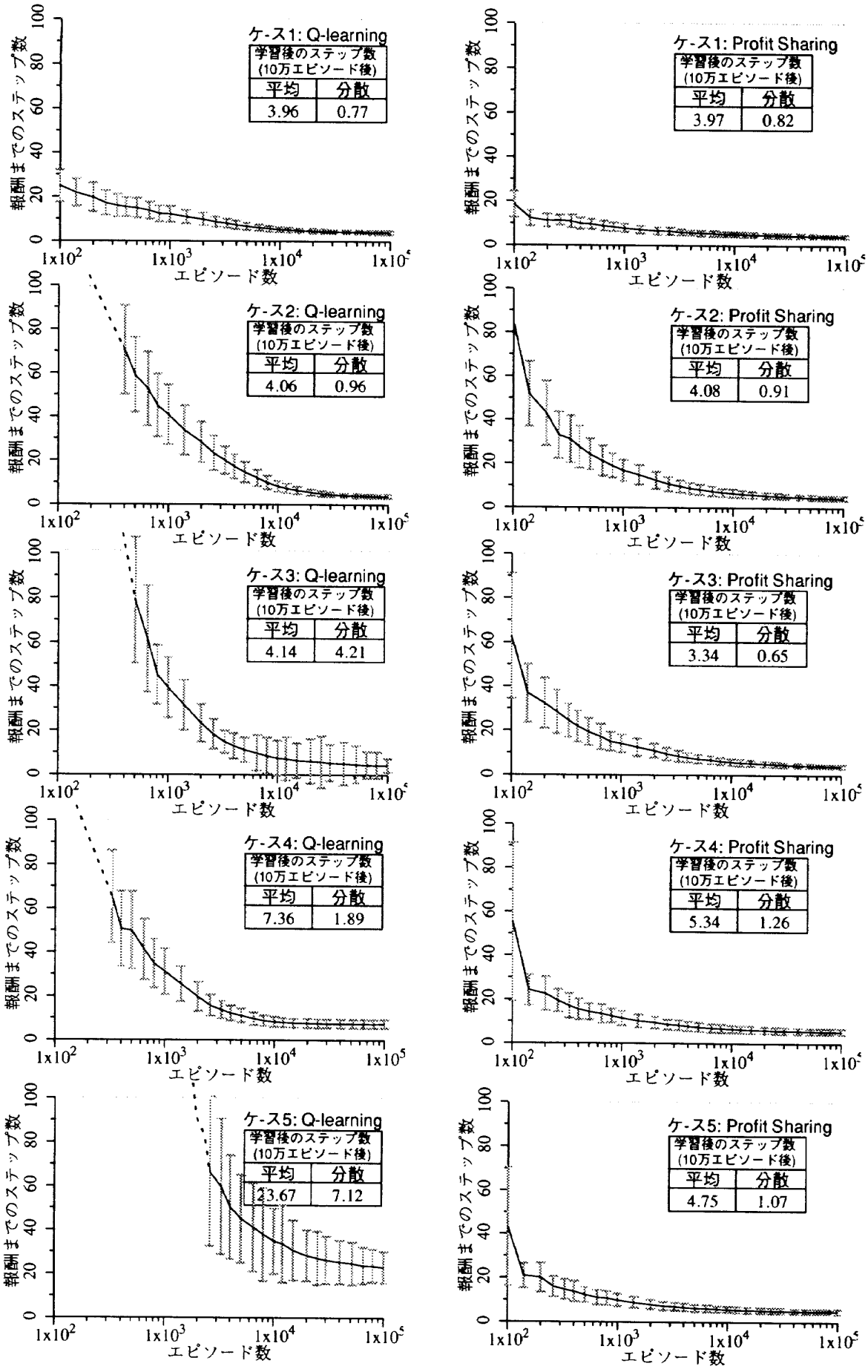


図4 追跡問題の各ケースにおける学習曲線

表1 Q-learningにおける学習定数αの影響

学習定数 α	学習後のステップ数(10万エピソード後)					
	ケース1		ケース2		ケース3	
	平均	分散	平均	分散	平均	分散
0.007	5.80	1.81	11.53	3.56	6.32	8.91
0.01	4.96	1.03	6.62	1.83	6.02	8.03
0.03	4.15	0.87	5.36	0.95	5.38	6.36
0.04	3.96	0.77	4.06	0.96	4.14	4.21
0.07	4.06	0.91	4.24	0.93	4.53	5.90
0.1	4.61	1.14	4.33	0.95	4.58	6.08
0.4	5.01	1.20	10.77	2.96	6.86	8.99
0.7	8.80	2.33	13.65	4.16	9.31	11.27

iii. ケース3の結果

ケース3では，獲物の逃避的行動がハンターの学習に影響を及ぼし，ケース2に比べ状態遷移の不確定性がより増大することから，特にQ-learningでは学習の立ち上がりが遅れるだけでなく，学習の後期でも分散が大きくなり，決定的な政策への収束を困難なものにしている．一方，Profit Sharingは状態遷移の不確定性の影響をあまり受けることなく，獲物の逃避行動に適応する政策を学習して，獲物を素早く捕獲することに成功している．

ケース2と3の結果より，他のハンターの学習および獲物の逃避的行動は状態遷移の不確定性を増大させ，これらはQ-learningの学習過程における環境同定を阻害する要因として作用し，学習を遅延させる結果につながっているといえる．

iv. Q-learningにおける学習定数αの影響

Q-learningでは，(1)式より，学習定数αの選択が学習性能に大きな影響を与えることから，ケース1～3においてαがどのような影響を及ぼすかを調べた．表1に10万エピソード学習後のエピソード長の10試行の平均と分散を示す．すべてのケースでαが0.04の場合が最も高い性能を示した．ケース1ではαに対し頑健であるのに対し，ケース2と3では頑健でなくなり，αが小さすぎても，大きすぎても平均と分散が大きくなっていることに注意されたい．これは，αが小さすぎるとBoltzmann選択がランダム選択に近づき，学習が遅くなる傾向があり，一方，αが大きすぎると学習が近視眼的となり，状態遷移の不確定性の影響をまともに受けてしまうためと考えられる．

〔2〕 **感覚入力が部分的な場合（ケース4,5）**

i. ケース4の結果

この場合，ケース2の状態遷移の不確定性に加えて，部分的な感覚入力に起因する不完全知覚が重なり，学習がより困難な問題になっている．Profit SharingのほうがQ-learningに比べて，不完全知覚の影響を受けにくく，学習後のエピソード長は2ステップ短く，より適切な政策を学習していることが分かる．Q-learningもそこそこの政策を学習できたのは，視界サイズに比べて環境サイズが小さいこと及び獲物の行動が非逃避的であることから，不完全知覚状態が少ないためと考えられる．

Profit Sharingが不完全知覚に対しても頑健な挙動を示すのは，不完全知覚状態に対しては有効な確率的政策が形成され，ルーレット選択によって不完全知覚状態から容易に脱出することが可能なためと考えられる．

ii. ケース5の結果

この場合，獲物の逃避的行動に伴う状態遷移の不確定性の増大と不完全知覚の相乗作用により，Q-learningにとっては最も学習が困難なケースとなっている．Profit Sharingとの比較で，Q-learningの性能悪化は歴然としている．

iii. 不完全知覚領域の拡大の影響

ケース4と5では，7×7の環境に対し，視界は5×5であり，不完全知覚領域は大きくはない．不完全知覚領域の拡大が学習の性能に及ぼす影響を調べるために，視界を5×5に固定して，環境が9×9及び15×15の場合について，同様の実験を行った．結果を図5に示す．ケース4では，環境が大きくなるにつれて，Q-learningとProfit Sharingではエピソード長の差異はより顕著となり，Q-learningが不完全知覚の影響をより強く受けていることが分かる．ケース5では，環境が9×9及び15×15の場合，Q-learningは収束せず，政策を形成することに失敗している．

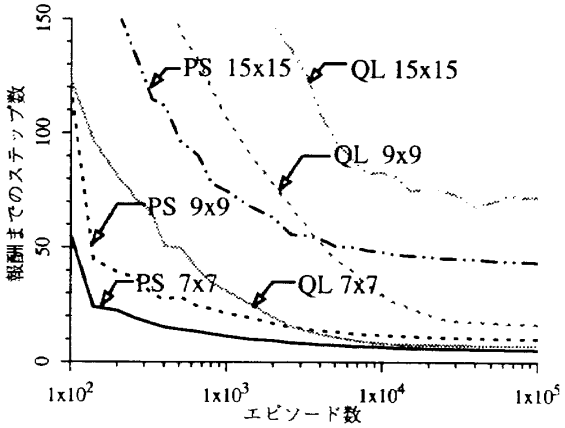
iv. 視界に何も見えない状態での行動の学習

不完全知覚領域の中で多数を占めるのは視界に何も見えない状態であり，環境サイズが大きくなるにつれて，その割合は増加する．例えば，視界5×5の2人のハンターがランダムに動く1匹の獲物を追ってランダムに動いた場合，7×7，9×9，15×15環境では，全状態に対し，それぞれ24%，49%，79%が何も見えない状態になる．したがって，何も見えない状態での行動の学習は非常に重要である．

ケース5の場合について，「何も見えない」状態でのQ値および重みに基づく行動選択の割合の変化を図6に示す．Q-learningは98 000～100 000エピソードの間の行動選択の割合の変化を，Profit Sharingは5 000エピソードまでの行動選択の割合の変化を示す．Q-learningは100 000エピソード付近でも行動選択の割合が振動を繰り返しているのに対し，Profit Sharingは5 000エピソード位で収束する傾向にあり，しかも2人のハンターが相補的な行動を強化して，意味のあ

学習後のステップ数(10万エピソード後)

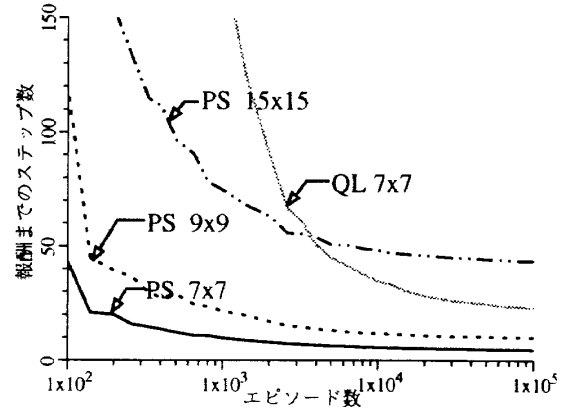
環境サイズ	学習方法			
	Q-learning		Profit Sharing	
	平均	分散	平均	分散
7×7	7.36	1.89	5.34	1.26
9×9	15.96	5.12	10.20	2.68
15×15	72.16	25.32	43.65	14.04



(a) ケース4の場合

学習後のステップ数(10万エピソード後)

環境サイズ	学習方法			
	Q-learning		Profit Sharing	
	平均	分散	平均	分散
7×7	23.67	7.12	4.75	1.07
9×9	収束せず		9.68	2.45
15×15			39.72	12.12



(b) ケース5の場合

図5 不完全知覚領域の拡大に伴うProfit SharingとQ-learningの挙動の変化

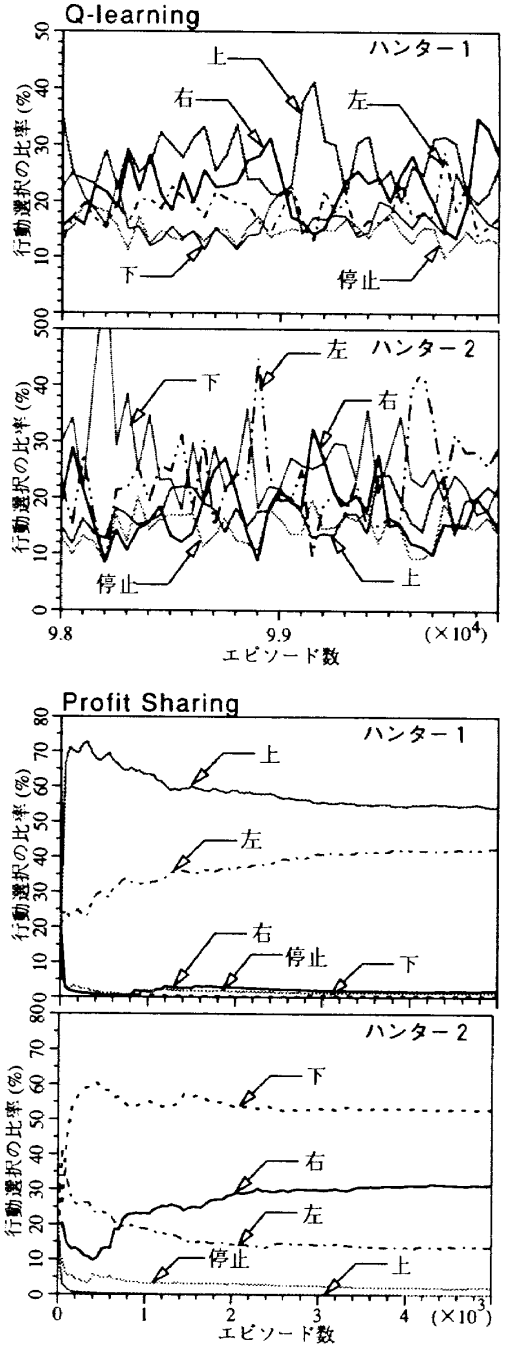


図6 ケース5において「視界内に何も見えない」状態における行動選択の比率の時間的変化

る協調的行動を取るための確率的政策を形成しつつあることが分かる．図6に示される違いが，Q-learningとProfit Sharingの間での性能の差を決定づけているといえる．

5. 結　　論

本論文は，マルチエージェント強化学習における問題の所在を明らかにした上で，Q-learningとProfit Sharingについて，追跡問題への適用を通じて，これらの手法の適用可能性を実験的に調べた．本論文で得られた成果は次のように要約される．

(1) マルチエージェント強化学習では，複数の学習エージェントの存在などに起因する状態遷移の不確定性，感覚入力の制限に起因する不完全知覚及び両者の相乗作用が学習を妨げる要因である．

(2) Q-learningは，状態遷移の不確定性が増大す

るにつれて，Profit Sharing に比べて学習の立ち上がり及び収束が遅れる傾向が強まるが，これはQ値の更新が式(1)，(2)に示されるような遷移先状態のQ値を用いるために，状態遷移の不確定性に対し敏感に反応し，Q値の変化が不規則になるためと考えられる．また，Q-learning が不完全知覚に対応できないのは，有効な確率的政策の形成が困難であり，不完全知覚状態からの脱出が容易ではないことによる．環境サイズが大きくなるにつれて，不完全知覚の影響をより強く受けるようになり，ケース5では9×9以上のサイズの環境では10万エピソード以上を費やしても収束しない．

(3) Profit Sharing が状態遷移の不確定性の影響を受けにくいのは，ルールの重みが割り引き報酬の累積値として記録されるので，状態遷移の不確定性を吸収する働きをもつためと考えられる．Profit Sharing が不完全知覚に対応できるのは，報酬フィードバックの繰り返しにより報酬に近づくための相補的なルールが獲得されること，及び確率的政策により不完全知覚状態から脱出できることによる．

(4) 本実験の結果より，各エージェントが独立に学習するマルチエージェント強化学習の枠組みとしてProfit Sharing は適切であるが，Q-learning はそのままでは適用することは困難である．

著者らは，現在，マルチエージェント強化学習の実問題(コイルヤードにおけるクレーンの群制御，エレベータの群制御など)へ適用を試みており，追跡問題と同様にProfit Sharing の有用性を確認している．

本論文では，Q-learning とProfit Sharing の比較に焦点をおいたので，ハンターが2人，環境が7×7の追跡問題を対象としたが，Profit Sharing はハンターが4人，環境が15×15以上の追跡問題においても有効に動作することを確認している．Profit Sharing は，規模が大きくなっても，立ち上がりが早く，安定した挙動を示しており，事前知識に依存することなく，協調的行動をボトムアップ的に創発する方法論として利用できると結論づけられる．Profit Sharing によって創発される協調的行動についてはたいへん興味深い結果が得られており，これは他の論文として公表する予定である．獲物が複数の場合，獲物に学習能力を持たせた場合，環境が動的に変化する場合などへ研究を展開することにより，マルチエージェント強化学習の本質を究めることが今後の課題である．

◇参考文献◇

[Gasser 89] Gasser, L., Rouquette, N., Hill, R.W. and Lieb, J.: Representing and Using Organizational Knowledge in Distributed AI Systems. in Gasser, L. and Huhns, M.H. (eds.),*Distributed Artificial Intelligence, Vol.2*, pp.55–78, Morgan Kaufmann (1989).

[Grefenstette 88] Grefenstette, J.J.: Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, *Machine Learning, Vol.3*, pp.225–245 (1988).

[Holland 86] Holland, J.H.:Escaping Brittleness:The Possibilities of General-Purpose Learning Algorithms Applied to Parallel Rule-Based Systems, in R.S.Michalsky, et al. (eds.), *Machine Learning: An Artificial Intelligence Approach, Vol.2*, pp.593–623, Morgan Kaufmann (1986).

[宮崎 94] 宮崎和光，山村雅幸，小林重信: 強化学習における報酬割当ての理論的考察， 人工知能学会誌, *Vol. 9, No. 4*, pp.580–587(1994).

[Ono 96] Ono, N., Ikeda, O. and Rahmani, A.T.: Synthesis of Herding and Specialized Behavior by Modular Q-learning Animats, *Proc. of the ALIFE V Poster Presentations*, pp.26–30 (1996).

[Sen 95] Sen, S. and Sekaran, M.: Multiagent Coordination with Learning Classifier Systems, in Weiss, G. and Sen, S.(eds.), *Adaption and Learning in Multi-agent systems*, Berlin, pp.218–233, Springer-Verlag, Heidelberg (1995).

[Singh 94] Singh, S.P., Jaakkola, T. and Jordan, M.I.: Learning Without State-Estimation in Partially Observable Markovian Decision Processes,*Proc. of the 11th International Conference on Machine Learning*, pp.284–292 (1994).

[Smith 81] Smith, R.G.: Framework for Cooperation in Distributed Problem Solving, *IEEE Trans. on Systems, Man and Cybernetics, SMC-11*, pp.61–70 (1981).

[Tan 93] Tan, M.: Multi-Agent Reinforcement Learning: Independent vs.Cooperative Agents, *Proc. of the 10th International Conference on Machine Learning*, pp.330–337 (1993).

[Watkins 92] Watkins, C.J.H., and Dayan, P.: Technical note: Q-learning, *Machine Learning, Vol.8*, pp.55–68(1992).

[Weiss 93] Weiss, G.: Learning to Coordinate Actions in Multi-Agent Systems, *Proc. of the 13th International Joint Conference on Artificial Intelligence*, pp.311–316 (1993).

[Weiss 95] Weiss, G.: Adaptation and Learning in Multi-Agent Systems: Some Remarks and a Bibliography, In Weiss, G. editor, *Lecture Notes in Artifitial Intelligence 1042*, pp.1–21, Springer-Verlag (1995).

[Werner 91] Werner, E.: The Design of Multi-Agent Systems, *Decentralized A.I. 3*, pp.3–30 (1991).

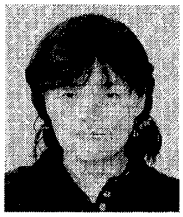
[Whitehead 90] Whitehead, S.D. and Balland, D.H.: Active perception and Reinforcement Learning, *Proc. of 7th International Conference on Machine Learning*, pp.162–169 (1990).

[Whitehead 91] Whitehead, S.D.: A Complexity Analysis of Cooperative Mechanisms in Reinforcement Learning, *Proc. of 9th National Conference on Artificial Intelligence, Vol.2*, pp.607–613(1991).

[山村 95] 山村雅幸，宮崎和光，小林重信: エージェントの学習， 人工知能学会誌, *Vol.10, No.5*, pp.683–689 (1995).

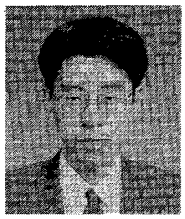
〔担当委員：安倍直樹〕

著者紹介



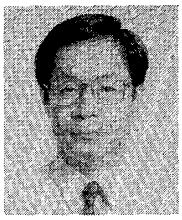
荒井　幸代(正会員)

1984年慶應義塾大学工学部計測工学科卒業．(株)ソニー，筑波大学大学院準研究員，米カリフォルニア大学バークレー校客員研究員を経て，1998年東京工業大学大学院総合理工学研究科知能科学専攻博士後期課程修了．工学博士．現在，同大学大学院総合理工学研究科リサーチアソシエイト．マルチエージェントの協調的行動に関する研究に従事．AAAI学会会員．　〈arai@fe.dis.titech.ac.jp〉



宮崎　和光(正会員)

1991年明治大学工学部精密工学科卒業．1996年東京工業大学大学院総合理工学研究科知能科学専攻博士後期課程修了．工学博士．同年4月，同大学大学院総合理工学研究科助手．1998年4月，同大学大学院総合理工学研究科リサーチアソシエイト．現在に至る．人工知能，特に強化学習に関する研究に従事．計測自動制御学会，日本機械学会各会員．　〈teru@fe.dis.titech.ac.jp〉



小林　重信(正会員)

1974年東京工業大学大学院博士課程経営工学専攻修了．同年4月，同大学工学部制御工学科助手，1981年8月，同大学大学院総合理工学研究科助教授，1990年8月，教授，現在に至る．知識システム，問題解決と推論制御，知識獲得と学習などの研究に従事．計測自動制御学会，情報処理学会各会員．　〈kobayashi@dis.titech.ac.jp〉