論　文

Technical Papers

Profit Sharingを用いたマルチエージェント強化学習における報酬配分の理論的考察

A Theory of Profit Sharing in Multi-Agent Reinforcement Learning

宮崎　和光*　荒井　幸代*　小林　重信*
Kazuteru Miyazaki　Sachiyo Arai　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.

1999年1月20日　受理

Keywords: multi-agent systems, reinforcement learning, profit sharing, rationality theorem, direct and indirect reward.

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.

In multi-agent reinforcement learning systems, it is important to share a reward among all agents. Conventional work has used ad hoc sharing schemes. We focus on the *Rationality Theorem* [Miyazaki 94] of Profit Sharing and analyze how to share a reward among all profit sharing agents. Though reward sharing may contribute to improve policy and learning speeds, it is possible to damage system behavior. It is important to preserve the *rationality condition* that expected reward per an action is larger than zero.

When an agent gets a *direct reward* R ($R > 0$), an *indirect reward* μR ($\mu \geq 0$) are given to the other agents. We have derived the necessary and sufficient condition to preserve the rationality condition as following;

$$\mu < \frac{M-1}{M^W\left(1-\left(\frac{1}{M}\right)^{W_0}\right)(n-1)L},$$

where M is the maximum number of conflicting rules in same sensory input, L is the maximum number of conflicting rational rules, W is the maximum episode length of direct reward agents, W_0 is reinforcement interval of indirect reward agents, and n is the number of agents.

This theory is derived by avoiding the least desirable situation whose expected reward per an action is zero. Theorefore, if we use this theorem, we can experience several efficient aspets of reward sharing. Through neumerical examples, we confirm the effectiveness of this theorem.

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，工学的およ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，従来DAI(Distributed Artificial Intellignece)のコミュニティを中心としたトップダウンアプローチが主であったが，近年，強化学習を用いたボトムアップ的手法が注目されている[荒井 98]．

マルチエージェント強化学習には，[荒井 98, Ono 96,

Sen 95, Tan 93, Weiss 93, Whitehead 91] など多くの研究がある．具体的方法として，[Ono 96, Tan 93, Whitehead 91] は Q-learning [Watkins 92] を，[Sen 95, Weiss 93] は Classifier System [Holland 86] をそれぞれベースにしている．また [荒井 98] は，Q-learning と Profit Sharing [Grefenstette 88] の二つの強化学習手法を取り上げ，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としての Profit Sharing の優位性を主張している．

現在多くのマルチエージェント強化学習研究においては，シングルエージェントを対象に開発された手法がそのまま利用されている．特に，直接報酬を獲得したエージェント以外のエージェントにいかに報酬を配分すべきかというマルチエージェント強化学習特有の問題に関しては，場当たりに的に解決されているに過ぎない．

一つの報酬をすべてのエージェントに配分することは，解の質および学習速度の向上につながる半面，システム全体の挙動に悪影響を及ぼす可能性もある．特に，負の報酬が存在しない場合には，システム全体の単位行動当たりの期待獲得報酬がゼロになるような配分は必ず避けなければならない．

本論文では，Profit Sharing に注目し，負の報酬が存在しない環境下で，システム全体の単位行動当たりの期待獲得報酬が正となるための各エージェントへの報酬配分に関する必要十分条件を導出する．これにより報酬が一切得られなくなるという最悪の状況を回避しつつ報酬配分による加速効果を受けることが可能となる．

以下，**2 章**では，マルチエージェント強化学習を定式化し，定理の導出に際して必要となる用語の定義ならびに仮定について述べる．**3 章**では，本論文の目的である**マルチエージェント系における合理性定理**を述べる．**4 章**では，定理の意味を理解するための各種数値例を述べる．**5 章**は結論であり，本研究の成果を総括し，今後の課題をとりまとめる．

2. 問 題 設 定

未知なる環境に置かれた n $(n>0)$ 個の機能的に均質なエージェント集団を考える．エージェントは環境からの**感覚入力**に対し，**行動**を選択し実行する．時間は，認識-行動サイクルによって離散化され，各時間単位には，n 個の中から任意に選ばれた一つのエージェントが一つの認識-行動サイクルを実行する．この認識-行動サイクルにおいて，ある特定のエージェントの

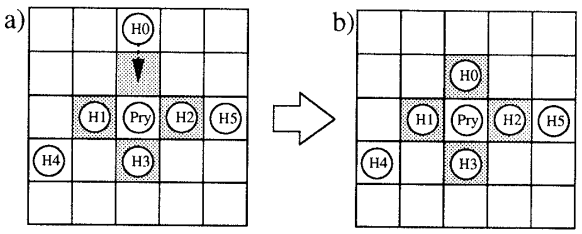


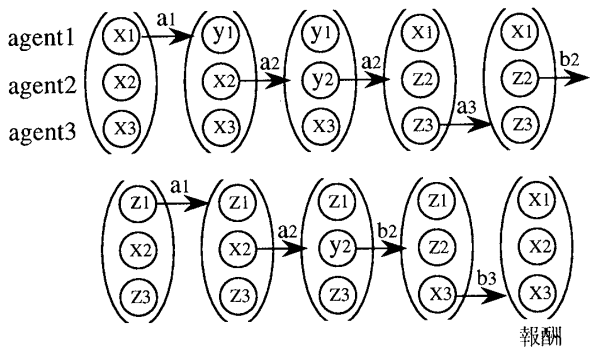
図 1 追跡問題の例．a) 報酬が与えられる直前の状態．b) 報酬が与えられる状態．

選択確率がゼロになることはないものとする．感覚入力は離散的な属性-値ベクトルである．エージェント i の感覚入力を $x_i, y_i, \cdots$ などと書く．行動は**競合数**と呼ばれる M 個の離散的なバリエーションの中から選ばれるものとする．エージェント i の行動を $a_i, b_i, \cdots$ などと書く．ある感覚入力において実行可能な行動は**ルール**として記述される．感覚入力 x で行動 a を選択する "if x then a" というルールを $\overrightarrow{xa}$ と書く．各感覚入力に対し，選択すべき行動を与える関数を**政策**と呼ぶ．単位行動当たりの期待獲得報酬が正である政策を**合理的政策**と呼び，それを最大化する政策を**最適政策**と呼ぶ．

ある時刻で $(n'-1)$ $(0<n'\leq n)$ 個のエージェントがそれぞれある特別な感覚を得ているときに，n' 個目のエージェントがある特別な感覚を得たとき，その n' 個目のエージェントに**直接報酬** R $(R>0)$ が与えられ，それ以外の $(n-1)$ 個のエージェントに**間接報酬** μR $(\mu \geq 0)$ が与えられるものとする．以下では，直接報酬を得るために必要な n' 個のエージェント群を**目標達成エージェント群**と呼ぶ．目標達成エージェント群は，組合せ的には nCn' 通り考えられる．

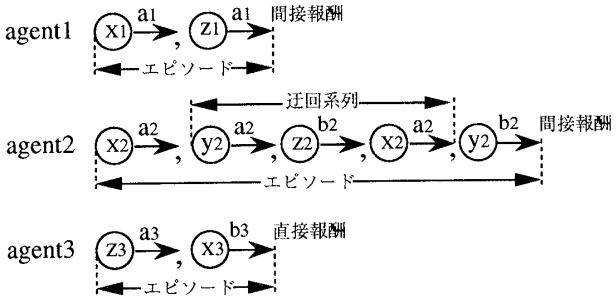
n' の値，および，ある特別の感覚が何であるかは一般には未知である．そのため，外界から報酬が与えられる直前に行動を出力したエージェントが報酬獲得に貢献したことは明らかであるが，それ以外のエージェントに関しては，報酬獲得に貢献したのか，妨害したのか，あるいは無関係なのかは一切わからない．したがって，μ を適切に設定することが非常に重要となる．

直接および間接報酬を図 1に示す追跡問題を例に説明する．この環境には，6 個のハンターと 1 個の獲物が存在する．今，ハンターが獲物の四方を取り囲んだ時点 (図 1b) で報酬が与えられるものとする．このとき，$n'=4$ であり，最後に獲物に接したハンターである H0 に直接報酬が与えられ，それ以外の 5 個のハンター (H1～5) に間接報酬が与えられる．n' の値および，ある特別な感覚が何であるかは未知なので，間接報酬が与えられるエージェント数は $n'-1=3$ ではなく $n-1=5$ となることに注意されたい．



Xi, Yi, Zi ; agent iの感覚入力　ai, bi ; agent iの行動

図 2　用語の説明に用いたルール系列



Xi, Yi, Zi ; agent iの感覚入力　ai, bi ; agent iの行動

図 3　エピソードおよび迂回系列の例

各エージェントごとに初期あるいは (直接または間接) 報酬を得た直後の感覚入力から次の報酬までのルール系列を**エピソード**という．例えば図 2に示すような $\overrightarrow{x_1a_1}\cdot\overrightarrow{x_2a_2}\cdot\overrightarrow{y_2a_2}\cdot\overrightarrow{z_3a_3}\cdot\overrightarrow{z_2b_2}\cdot\overrightarrow{z_1a_1}\cdot\overrightarrow{x_2a_2}\cdot\overrightarrow{y_2b_2}\cdot\overrightarrow{x_3b_3}$ というルール系列が与えられた場合，この中にはエージェント 1,2,3 ごとに，それぞれ $(\overrightarrow{x_1a_1}\cdot\overrightarrow{z_1a_1})$, $(\overrightarrow{x_2a_2}\cdot\overrightarrow{y_2a_2}\cdot\overrightarrow{z_2b_2}\cdot\overrightarrow{x_2a_2}\cdot\overrightarrow{y_2b_2})$, $(\overrightarrow{z_3a_3}\cdot\overrightarrow{x_3b_3})$ というエピソードが含まれている (図 3参照)．以下では，直接報酬獲得時に形成されるエピソードの開始点はつねに同一であるものとする．

あるエピソードで，同一の感覚入力に対して異なるルールが選択されているとき，その間のルール系列を**迂回系列**という．例えばエージェント 2 のエピソード $(\overrightarrow{x_2a_2}\cdot\overrightarrow{y_2a_2}\cdot\overrightarrow{z_2b_2}\cdot\overrightarrow{x_2a_2}\cdot\overrightarrow{y_2b_2})$ には，迂回系列 $(\overrightarrow{y_2a_2}\cdot\overrightarrow{z_2b_2}\cdot\overrightarrow{x_2a_2})$ がある (図 3参照)．迂回系列上のルールは，報酬の獲得には貢献しない可能性がある．現在までのすべてのエピソードで，常に迂回系列上にあるルールを**無効ルール**と呼び，それ以外を**有効ルール**と呼ぶ．例えば図 2に示すルール系列を得た後には，$\overrightarrow{y_2a_2}$ および $\overrightarrow{z_2b_2}$ が無効ルールであり，それ以外のルールが有効ルールである．

本論文では Profit Sharing (PS) の使用を前提とする．PS ではエピソード単位でルールの評価値としての重みを強化する．**強化関数**とは，報酬からどれだけ過去かを引き数とし，**強化値**を返す関数である．f_i によって報酬から i ステップ前の強化値を参照する．長さ W_a のエピソード $(r_{W_a-1}\cdots r_i\cdots r_1\cdot r_0)$ に対して，ここでは $S_{r_i}=S_{r_i}+f_i$ によってルールの重み (S_{r_i}) を強化する場合について考える．ここで，W_a を**強化区間**と呼ぶ．

直接報酬を獲得したエージェントが，R $(R>0)$ の報酬を得たとき，強化関数には，行動のバリエーションが M 個の場合の**シングルエージェント系における合理性定理** [宮崎 94] を満たす (1) 式のような等比減少関数を用いる．

$$f_n=\frac{1}{M}f_{n-1},\quad n=1,2,\cdots,W_a-1. \qquad (1)$$

ここで，初期報酬値 f_0 および強化区間 W_a は，

直接報酬を獲得したエージェントでは

$$f_0=R,\quad W_a=W$$

それ以外のエージェントでは

$$f_0=\mu R\ (\mu\geq 0),\quad W_a=W_0\ (W_0\leq W)$$

とする．

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においては，複数エージェントの同時学習によって生じる**状態遷移の不確定性**，およ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** [Whitehead 91] の二つの問題が重要である [荒井 98]．さらにこれらの問題は，[宮崎 99] における**タイプ 1 の混同**および**タイプ 2 の混同**という二つの観点から捉え直すこともできる．ここでタイプ 1 の混同とは，価値の高い状態と価値の低い状態が同一視されることをいい，タイプ 2 の混同とは，あるとき有効ルールであると判定されたルールが，ある時点以降，つねに迂回系列上に存在してしまうことをいう．一般に，タイプ 2 の混同が存在すれば，タイプ 1 の混同も同時に存在する．タイプ 1 およびタイプ 2 の混同という観点から環境は図 4に示すような三つのクラスに分類される．ここで Q-learning (QL) などで最適性が保証されているマルコフ決定過程の環境はクラス 1 に属する．

QL に代表される動的計画法に基づく手法は，タイプ 1 の混同の影響を強く受けるが，PS は状態の価値を利用しないので，タイプ 1 の混同に対し頑健である [荒井 98, 宮崎 99]．一方，タイプ 2 の混同に関しては，PS,QL ともにその影響を強く受ける．本論文では PS の使用を前提としているので，証明上，タイプ 2 の混同が存在しないクラスを対象とする必要がある．

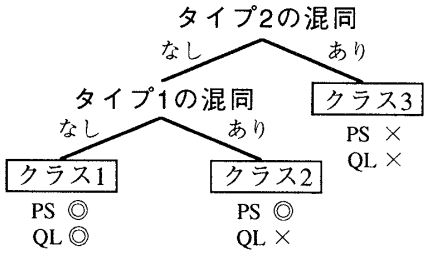


図4 マルチエージェント強化学習における困難さの分類.

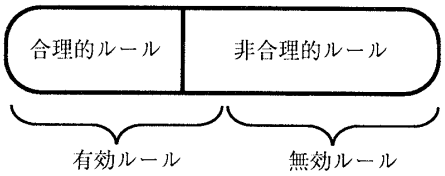


図5 ルールの種類.

しかし実際上は，[荒井 98] に見られるように，確率的政策によりタイプ 2 の混同からの脱出は十分可能である．また，この仮定は，図 4からもわかる通り，マルチエージェント強化学習の特徴である状態遷移の不確定性や不完全知覚のすべてを排除するものではない．例えば，ひとたび遷移した経験のある遷移先への状態遷移確率がゼロにならなければ，たとえ状態遷移の不確定性や不完全知覚が存在していたとしてもタイプ 2 の混同は存在しない．よって本論文が対象とするクラスは，マルチエージェント強化学習として十分意味のあるクラスであると考える．

[宮崎 94] の合理性定理を満たす$\mu = 0$ の PS で報酬が得られるとき，その PS によって獲得され得る有効ルールを**合理的ルール**，それ以外のルールを**非合理的ルール**と呼ぶ．これらのルールの包含関係を図 5に示す．非合理的ルールと合理的ルールとが競合するならば，明らかに非合理的ルールを強化すべきではない．$\mu > 0$ の場合，一部の非合理的ルールが有効ルールとみなされる可能性がある．したがって，学習システムには，有効ルールの中から，非合理的ルールを排除する機能が求められる．

次章では，上記の観点から，システム全体の単位行動当たりの期待獲得報酬を正とするためのμの範囲に関する必要十分条件を導出する．

3. Profit Sharing を用いたマルチエージェント強化学習における合理性定理

3・1 基 本 方 針

システム全体の単位行動当たりの期待獲得報酬を正とするためには，非合理的ルールが抑制されたエージェント群により目標達成エージェント群が構成される必要がある．ここであるエージェント群において非合理的ルールが抑制されるとは，そのエージェント群内の各エージェントにおいて，非合理的ルールがそれと競合する合理的ルールを差し置いて一番には強化されないことである．また，逆に，非合理的ルールが抑制されたエージェント群により目標達成エージェント群が構成されていれば，それらのエージェント各々に関してはつねに合理的ルールの選択が可能となり，システム全体の単位行動当たりの期待獲得報酬は正となる．

よって，本章では，等比減少関数を用いた PS において，ある目標達成エージェント群における非合理的ルールを抑制するためのμの範囲に関する必要十分条件を求める．

以下では，まず，非合理的ルールを抑制することが最も困難となるルールの競合構造を選ぶ．ここで二つの競合構造 A と B について，A において非合理的ルールを抑制できるμの範囲が B のそれに包含されるとき，**A は B よりも困難**であるという．次に，最も困難な構造に対して，非合理的ルールを抑制するためのμの範囲に関する必要十分条件を求める．最後にそれを任意のルールの競合構造に対して拡張する．

3・2 マルチエージェント強化学習における非合理的ルール抑制定理

［補題 1］（最も困難な構造）
唯一の回帰的非合理的ルールの抑制が最も困難である．
□

証明は**付録 A** に示す．図 6に最も困難な競合構造を示す．L 本の合理的ルールと唯一の回帰的非合理的ルールが競合している．ここで，行動をとった結果，感覚入力の変化が生じないルールは**回帰的**であるという．

［補題 2］（唯一の回帰的非合理的ルールの抑制）
ある目標達成エージェント群において唯一の回帰的非合理的ルールが抑制されるためのμの範囲に関する必要十分条件は

$$\mu < \frac{M-1}{M^W(1-(\frac{1}{M})^{W_0})(n-1)L}. \qquad (2)$$

□

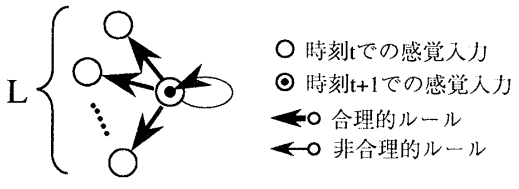


図6 最も困難な競合構造.

証明は**付録B**に示す．ここで，Wは直接報酬を獲得したエージェントの最大エピソード長，W_0は直接報酬を獲得したエージェント以外のエージェントの強化区間，Mは競合数，Lは同一感覚入力下に存在する有効ルールの最大競合数である．

補題1と2から推移律により直ちに次の定理が得られる．

［定理］（マルチエージェント系における合理性定理）ある目標達成エージェント群において任意の非合理的ルールが抑制されるためのμの範囲に関する必要十分条件は(2)式である．

3・3 実装上のポイント

一般にLの値は学習以前には知ることはできない．しかしLの値が最大となるのは，唯一の無効ルールと競合した場合なので，実装上は$L = M - 1$とすれば十分である．

一般にWの値は学習以前には知ることはできないが，実装上はWを任意の値に設定し，ある時点でのエピソード長がWを超えた場合には$\mu = 0$とし，一切の間接報酬を与えなければ十分である．

本論文では，直接報酬により形成されたエピソードの開始点はつねに同一であるとしたが，2種類以上の開始点が存在した場合には，ある一定の開始点以外を開始点とした際は，$\mu = 0$とする等の工夫を導入すればよい．

$L = M - 1$, さらに$W_0 = W$とし，間接報酬による強化を直接報酬による強化と同じ強化区間まで許した場合，(2)式は以下のように簡単化される．

$$\mu < \frac{1}{(M^W - 1)(n - 1)}. \qquad (3)$$

3・4 定理の意味

定理1は，間接報酬が非合理的ルールを強化してしまう場合，すなわち，間接報酬による副作用を抑制するために導出された定理である．一般には，間接報酬が合理的ルールを強化するような加速効果も十分考えられる．定理1は，そのような加速効果に関しては何も保証していないが，報酬が全く得られなくなるという最悪の事態を回避した上で，学習システムの利用者に，間接報酬による加速効果を期待させるものとして非常に意味がある．

仮に，各エージェントがある一定の順番で認識-行動サイクルを繰り返すならば，全体を一つのエージェントとして捉え，[宮崎 94]のシングルエージェント系における合理性定理により学習させることも可能である．しかし定理1は，各エージェントがランダムに認識-行動サイクルを繰り返す場合にも成り立つ．したがって本定理は，[宮崎 94]の定理よりもより広いクラスに対し，合理性を保証することができる．

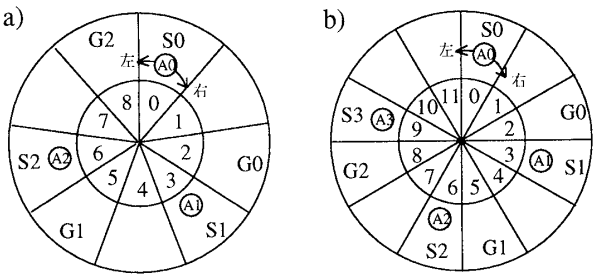
4. 数値例

4・1 実験環境

図7に示す2種類のルーレット状の環境で実験を行った．環境aには3, 環境bには4個のエージェントが存在する．エージェントi (A_i)の初期位置がS_iである．各エージェントへは，各図の中央部分に示したルーレットの目が感覚入力として与えられる．各エージェントの行動集合は，ルーレットの中央を原点として，右隣の目に移動，もしくは，左隣の目に移動の2種類である．右への行動は20%, 左への行動は50%の確率で失敗し，失敗した場合，現在のます目にとどまる．エージェント同士は重なることはできない．

あるエージェントiがゴールG_iに存在し，かつ，エージェントR $(R \neq i)$がゴールG_Rに到達したとき，エージェントRに直接報酬R $(= 100.0)$が与えられ，それ以外のエージェントに間接報酬μRが与えられる．あるエージェントが直接報酬を獲得するか，または，あるエージェントiがゴールG_j $(j \neq i)$に到達したときに，すべてのエージェントが，図7に示す初期配置に戻される．環境a)ではすべてのエージェントがゴールを持つが，環境b)ではエージェント0はゴールを持たない．したがって，環境b)は環境a)に比べ，間接報酬の重要性が高い問題であると言える．

認識-行動サイクルは，環境a)では，エージェント0〜2が，それぞれ，0.9, 0.05, 0.05の確率で選ばれ，環境b)では，エージェント0〜3が，それぞれ，0.72, 0.04, 0.04, 0.2の確率で選ばれることで繰り返される．エージェントiの認識-行動サイクルは，エー



AiがGj (j=i)に到達 → 報酬＆初期化
AiがGj (j≠i)に到達 → 初期化

図7 数値例で用いたルーレット状環境．

ジェント iがゴール G_iに到達した時点で停止し，あるエージェントが直接報酬を獲得した時点で再開される．各認識-行動サイクルにおいて,「感覚入力，行動選択実行，状態遷移，報酬判定」の4ステップが順次実行される．

環境a),b)ともに，すべてのエージェントが「つねに右」という行動を出力する政策が単位行動当たりの期待獲得報酬を最大化する**最適政策**である．また，エージェント0〜2の中の少なくとも二つが「初期位置で左」もしくは「初期位置で右かつ初期位置の右隣で左」という行動を出力する政策が単位行動当たりの期待獲得報酬をゼロにする**非合理的政策**である．

4・2 実験条件

直接報酬を獲得したエージェントの最大エピソード長 Wは学習以前には不明だが，**3・3節**に従い，実装上 $W=3$とし，ある時点でのエピソード長が3を越えたとき$\mu=0$とした．$W=3$のとき(3)式より，環境aでは$\mu<0.0714...$，環境bでは$\mu<0.0333...$の範囲で定理を満たす．

学習の初期段階では，獲得された政策が振動する可能性が高いので，最適政策に関しては，それが形成された時点から，100エピソードの間維持された場合に限り，学習成功と判定する．学習の打ち切りは，エージェント0〜2が共に「初期位置で左」という非合理的政策を獲得した時点，または，全行動数が10万に達した時点とする．各ルールの初期重みは100.0である．

4・3 結果および考察

図7に示した2種類の環境に対する，学習の結果得られた政策の質ならびに1000回の最適政策の学習に要した行動数の平均と標準偏差をそれぞれ表1, 表2に示す．ここで，政策の質は，乱数の種を変えて行った1000回の実験中の非合理的および最適政策の獲得回数で評価している．さらに，図8, 図9に，1000回の最適政策の学習に要した行動数の平均と標準偏差の詳細を示す．

定理は単位行動当たりの期待獲得報酬を正とする合理性のみを保証しており，最適性までは保証していないが，環境a),b)ともに定理の範囲を多少越えた部分においても，つねに最適政策が獲得された．定理は，すべての環境に対し合理性を保証するものなので，このような比較的簡単な環境に対しては，より緩い条件で合理性や最適性が満たされる．

環境a)では，表1および図8より，$\mu=0.3$の場合

表1 環境a)における，学習の結果得られた政策の質ならびに1000回の最適政策の学習に要した行動数の平均と標準偏差．

μ	学習終了後の政策獲得回数		最適政策の学習に要した行動数	
	非合理	最適	平均	標準偏差
0.0	0	1000	1201.1	273.0
0.000001	0	1000	1031.2	119.4
0.07	0	1000	946.7	107.0
0.3	**0**	**1000**	**900.7**	**172.8**
0.4	1	999	910.3	221.3
1.0	4	939	1120.0	794.3

表2 環境b)における，学習の結果得られた政策の質ならびに1000回の最適政策の学習に要した行動数の平均と標準偏差．

μ	学習終了後の政策獲得回数		最適政策の学習に要した行動数	
	非合理	最適	平均	標準偏差
0.0	0	0	−	−
0.000001	0	1000	2690.2	263.8
0.03	0	1000	2570.6	265.2
0.2	**0**	**1000**	**2474.9**	**402.6**
0.4	1	998	2671.8	945.2
1.0	13	909	3103.8	1561.6

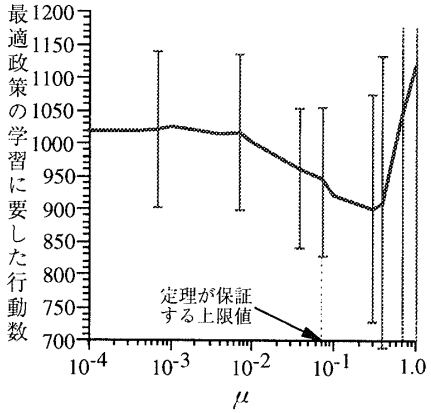


図8 環境a)における，1000回の最適政策の学習に要した行動数の平均と標準偏差の詳細．

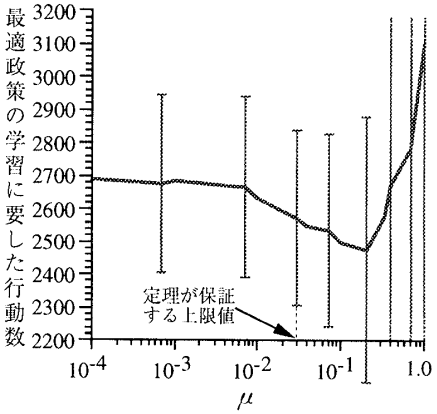


図9 環境b)における，1000回の最適政策の学習に要した行動数の平均と標準偏差の詳細．

が，最も少ない行動数で最適政策を獲得している．一方0.4以上では，非合理的政策が獲得される場合がある．例えば，エージェント1と2がG_0およびG_1に接近する行動が間接報酬により一番に強化されてしまい非合理的政策となる．しかし定理を満たすμの場合，そのような誤った強化は決して起こらない．ここでみられるように，μが定理の範囲内であれば，結果として直接報酬を獲得したエージェント以外のエージェントの合理的ルールが強化され，学習速度の向上につながる場合は多い．さらに，定理の範囲を越えても学習速度が向上する可能性もあるが，すべての環境に対し合理性を保証するという観点からは，定理の範囲内での最大値にとどめておくことが望ましい．

環境b)では，エージェント3は，ゴールを持たないため，それ自身では学習不可能である．したがって，$\mu = 0$の場合，最適政策を獲得することができない．このような場合，間接報酬が威力を発揮する．表2および図9より，$\mu = 0.2$の場合が，最も少ない行動数で最適政策を獲得しており，また，$\mu \leq 0.3$の範囲であれば，つねに最適政策が獲得されている．この環境のように，$\mu = 0$よりも$\mu > 0$とした場合の方が，解の質が向上することは，間接報酬の一つの重要な利点である．

定理は，学習システムとしての最低限の条件，すなわち，システム全体の単位行動当たりの期待獲得報酬を正とするためのものであるが，数値例により解の質や学習速度の向上にも多いに貢献する可能性があることが示唆された．

5. おわりに

マルチエージェント強化学習研究は数多いが，現在までのところ，単に，シングルエージェントを対象に開発された手法をそのまま利用しているに過ぎない．特に，直接報酬を獲得したエージェント以外のエージェントにいかに間接報酬を配分すべきかというマルチエージェント強化学習特有の問題に関しては，経験的知見の域を越えていない．

本論文では，Profit Sharingに注目し，システム全体の単位行動当たりの期待獲得報酬を正とするための各エージェントへの報酬配分に関する必要十分条件を導出した．多くのエージェントに報酬を配分すれば，配分しない場合に比べ，解の質および学習速度の向上が期待できる半面，システム全体の挙動に悪影響を及ぼす可能性がある．しかし本論文で示された条件を用いれば，報酬が得られなくなるという最悪の事態を回避しつつ，報酬配分の恩恵を受けることができる．

今回行った解析は，間接報酬による副作用の抑制に着目したものである．一般には，間接報酬が有効に働く場合も多く，今後は，間接報酬による加速効果を早急に解析すべきと考える．また，定理を満たす範囲内でのより強力かつ有効な条件，工学的応用なども今後の重要な課題である．

◇参考文献◇

[荒井 98] 荒井幸代，宮崎和光，小林重信: 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—，人工知能学会誌，Vol. 13, No. 5, pp. 609–618 (1998).

[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, Morgan Kaufmann, pp. 55–78(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 Sysems, in R.S. Michalsky et al. (eds.), *Machine Learning: An Artificial Intelligence Approach*, Vol. 2, pp. 593–623. Morgan Kaufman (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).

[宮崎 99] 宮崎和光，荒井幸代，小林重信：POMDPs環境下での決定的政策の学習，人工知能学会誌，Vol. 14, No. 1, pp. 148–156 (1999).

[Ono 96] Ono, N., Ikeda, O. and Rahmani, A.T. : Synthesis of Herding and Specialized Behavior by Modular Q-learning Animates, *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, Heidelberg. Springer Verlag, pp. 218–233 (1995).

[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).

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

[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).

〔担当委員：阿久津達也〕

◇ 付　　録 ◇

A. 補題1の証明

簡単のため $L=1$ とする．それ以外も全く同様である．

明らかに，非合理的ルールが強化される回数が多いほど，非合理的ルールを抑制できるμの範囲は狭くなり，非合理的ルールの抑制はより困難となる．よって，強化される回数の大小のみを考えれば十分である．**枝分かれ数**の小さい順に可能な競合構造を数え上げる．ここで枝分かれ数とは，可能な状態遷移の総数である．

1) 枝分かれ数が1の場合

競合数1なので，明らかに困難ではない (図 A.1-a)．

2) 枝分かれ数が2の場合

競合数が1ならば 1) と同様 (図 A.1-b)．競合数が2の場合を考える．回帰ルールを含む場合 (図 A.1-c) とそうでない場合 (図 A.1-d) とに分けられる．ここで A を合理的ルール，B を非合理的ルールとする．任意のエピソードで，1回の A につき B は繰り返し使われた可能性がある．図 A.1-c の方が図 A.1-d よりも1回の A につき B を選ぶ回数を多くできる．したがって，図 A.1-d よりも図 A.1-c の方がより困難な構造である．

3) 枝分かれ数が3の場合

競合数が1の場合は 1) と同様 (図 A.1-e)．競合数が2の場合 (図 A.1-f) は 2) と同様に，唯一の回帰的非合理的ルールと競合する場合が最も困難となる．競合数が3の場合 (図 A.1-g) 非合理的ルールが全部で2個なので，非合理的ルール1個あたりの強化回数は1個の場合よりも減少する．よって，競合数2のときが最も困難である．

同様に，任意の枝分かれ数のときも，唯一の回帰的非合理的ルールと競合する場合が最も困難となる．

ゆえに，非合理的ルールを抑制することが最も困難となる構造は，合理的ルールが唯一の回帰的非合理的ルールと競合する構造である．　　(証明終り)

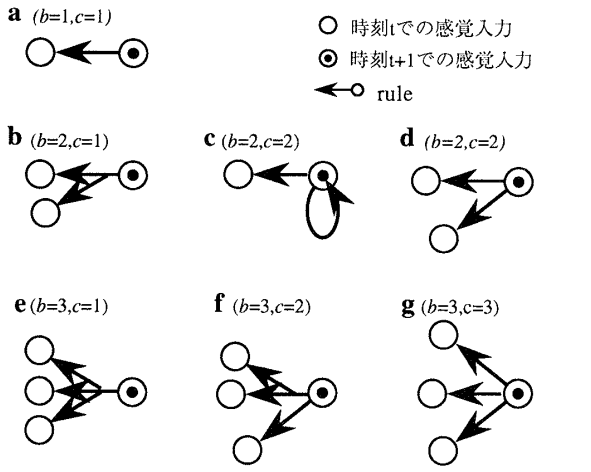


図 A.1　証明で用いた競合構造．

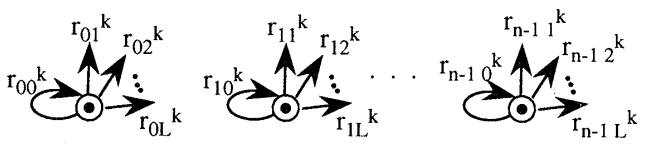


図 A.2　証明で用いたマルチエージェント構造．

B. 補題2の証明

ある目標達成エージェント群の任意の強化区間 k ($k=0,1,\ldots,W-1$) に対し，

$$S_{ij}^k > S_{i0}^k \tag{B.1}$$

となる j ($=1,2,\ldots,L$) が存在することを示す．ここで S_{ij}^k はエージェント i の j 番目の合理的ルール (r_{ij}^k) の重み，S_{i0}^k はエージェント i の唯一の回帰的非合理的ルール (r_{i0}^k) の重みである (図 A.2参照)．

まず r_{ij}^k と r_{i0}^k の選択回数の比を考える．r_{ij}^k と r_{i0}^k の選択回数の比の最小値が最大となるのは目標達成エージェント群の個数 n' が1の場合であり，かつ，各エージェントにおける L 本の合理的ルールが各エージェントごとに順番に選択された場合である (表 B.1 参照)．このとき，

$$r_{ij}^k\text{の選択回数} : r_{i0}^k\text{の選択回数} = 1:(n-1)L \tag{B.2}$$

となる (図 B.3参照)．

次に r_{ij}^k と r_{i0}^k 各々に与えられる強化値について考える．r_{ij}^k に与えられる強化値が最小となるのは，直接報酬を獲得したエージェントが W 区間すべてにおいて異なる状態を知覚していた場合

表 B.1　ある k における全エージェントに関するルールの選択系列．例えば，この表で，1ヶ所でも非→合$_1$や合$_2$→合$_1$となれば，その合$_1$を選択可能とするエージェントにおける学習はより容易になる．

報酬獲得回数	エージェント番号 0	1	…	n-1
	合1	非	…	非
	合2	非	…	非
	⋮			
	合L	非	…	非
	非	合1	…	非
	非	合2	…	非
	⋮			
	非	合L	…	非
	⋮			
	非	非	…	合1
	非	非	…	合2
	⋮			
	非	非	…	合L

合；合理的ルール
非；非合理的ルール

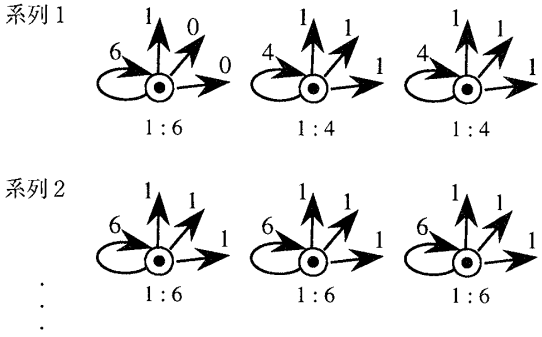


図 B.3　ルールの選択系列の例．系列1のように各ルールの選択に偏りがある場合の学習は比較的容易．系列2のように均等に各ルールが選択された場合が最も困難．

であり，特に $k = W$のとき最小値 $R/(M^{W-1})$ をとる．一方，r_{i0}^{k}に与えられる強化値が最大となるのは，直接報酬を獲得したエージェント以外のエージェントがW区間すべてにおいて同一の状態を知覚していた場合であり，$W \geq W_0$の範囲で

$$\mu R \frac{M}{M-1}\left(1-\left(\frac{1}{M}\right)^{W_0}\right)$$

をとる．

よって (B.1) 式が成立するためには，

$$\frac{R}{M^{W-1}} > \mu R \frac{M}{M-1}\left(1-\left(\frac{1}{M}\right)^{W_0}\right)(n-1)L. \tag{B.3}$$

すなわち

$$\mu < \frac{M-1}{M^{W}\left(1-\left(\frac{1}{M}\right)^{W_0}\right)(n-1)L} \tag{B.4}$$

が必要である．十分性は明らかである．　(証明終り)

著　者　紹　介

宮崎　和光(正会員)は，前掲（Vol. 14, No. 1, p. 156）参照．

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．ソニー（株），米 UC.Berkeley 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，現在，米 Carnegie Mellon University, Research Associate. マルチエージェントシステム研究に従事．AAAI 学会会員．　〈arai@fe.dis.titech.ac.jp〉

小林　重信(正会員)は，前掲（Vol. 14, No. 1, p. 130）参照．