強化学習による協調的行動の創発に関する実験的考察

東京工業大学 総合理工学研究科　○ 白川 英隆，木村 元，小林重信

Experimental study on Emergence of cooperative action using Reinforcement Learning

○ Hidetaka SHIRAKAWA,Hajime KIMURA, Shigenobu KOBAYASHI

Graduate School of Interdisciplinary Science and Eng.,

TOKYO INSTITUTE OF TECHNOLOGY

Abstract: Multiagent systems in which agents interact with each other are now being proposed as a solution to many problems which can be grouped together under the "distributed problem solving". For such system to work properly, it is necessary that agents learn from their environment and other agents and adapt their behaviour. Though the reward is important information of the multiagent system,no research studied about it.

We propose a method of reward-distribution and provide experimentl studies. And We apply the Stochastic Gradient Ascent and suggest that it can be a base of theory on multiagent reinforcement learning.

1　はじめに

今日，高度に複雑化，巨大化したシステムをトップダウン的に設計することには限界があると言われ，各構成要素からボトムアップ的に設計する枠組の一つとして，マルチエージェントシステムが注目を浴びている．近年では特に，各構成要素についても自律的に行動を学習する必要があるため，行動の結果として得られる賞や罰を報酬として行動の学習を行う強化学習を用いた，マルチエージェント強化学習についての研究が行われている．

マルチエージェント強化学習の従来研究としては，主に 2 つの手法として Q-learning[Watkins 92] を用いた研究 [Tan 93]，[Ono 96] と Profit Sharing を用いた研究 [Arai 97] がある．近年研究が進みつつあるマルチエージェント強化学習ではあるが問題点もいくつかある．まず，強化学習における報酬は，学習の目的を表す唯一の情報であり，マルチエージェントシステムにおいては，全体の報酬を考慮しなければならないにも関わらず，考察はほとんどされてきていない．また，シングルエージェントの拡張として強化学習を用いて実験的に試行されている研究は多いが，理論的な考察を加えた研究はほとんどない．

そこで，本研究ではマルチエージェント強化学習において，報酬をどのようにエージェントに与えたら良いかという視点から，報酬の配分方法に関しての実験的考察を行う．また，強化学習法としては確率的傾斜法 [Kimura 95] を適用し，報酬を全てのエージェントが共有したときにおける理論的な考察 [木村 97] についても簡単に触れ，マルチエージェント強化学習における報酬全般に関して議論する．

以下 2 章では報酬の配分方法を扱う問題設定について述べる．3 章では扱うマルチエージェントの対象として搬送問題を単純に抽象化し，報酬の配分方法を変化させ実験を行ない，配分方法による特徴を明らかにする．4 章では本研究の成果をまとめる．

2　問題設定

2.1　従来のマルチエージェント強化学習研究

マルチエージェント強化学習は行動主体であるエージェントが複数存在する図 1のような環境で，各々のエージェントは学習器を持ちマルチエージェントシステムから与えられる報酬により学習を行う．

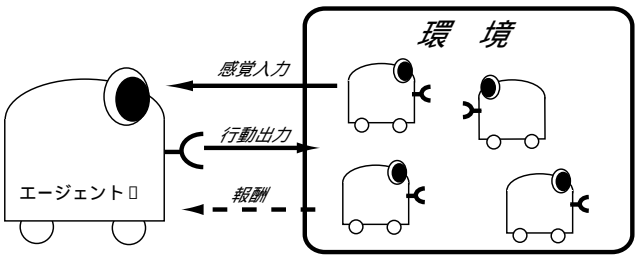


図 1: マルチエージェント環境

マルチエージェント強化学習に関する従来研究としては，主に 2 つの接近法が研究されている．1 つは環境同定型の代表的な強化学習である Q-learning[Watkins 92] をベースとした学習器を用いる接近法 [Tan 93][Ono 96] で，もう 1 つは経験強化型の強化学習である Profit Sharing をベースとした学習器を用いる接近法 [Arai 97] である．どちらの手法もシングルエージェントで用いた強化学習アルゴリズムをマルチエージェント強化学習に適用した研究であると位置付けられ，実験的に研究が行なわれて来ているが，理論的に解析はほとんど行なわれていない．

2.2　確率的傾斜法 (Stochastic Gradient Ascent:SGA)

本研究で用いる確率的傾斜法について説明する．

確率的傾斜法は Dynamic Programming に基づかない経験強化型の強化学習である．不完全知覚により部分的な観測

を得るマルコフ決定過程を部分観測マルコフ決定過程 (Partially Observable Markov Decision Process:POMDP) と呼び，確率的傾斜法は POMDP でモデル化される環境に対して，平均報酬を最も大きくする方向へ逐次的に政策を改善する手法である．

確率的傾斜法は適正度 (eligibility)[Williams 92] と呼ばれる，どんな行動を取ったのかについての情報について計算し，過去の行動ほど割り引き率γで減衰 / 忘却させながら適正度の履歴 (eligibility trace) として保存してある．適正度の履歴にしたがって重みを更新するため，報酬に遅れのある行動も強化される．

具体的には観測入力 X に対して行動 a を選択する確率を政策πと呼び，関数$\pi(a, W, X)$ で表す．Wは一般的にエージェント内部の変数ベクトルで，これを調節することで政策を改善し学習を行う．ニューラルネットならば Wはリンクの重みにあたり，重みつきルールベースシステムなら W はルールの重みに相当する．

図 2に SGA による強化学習の一般形を示す．

1. 観測 $X(t)$ を受け取る
2. 確率$\pi(a(t), W, X(t))$ で，行動 $a(t)$ を実行
3. 報酬 $r(t)$ を受け取る
4. 適正度 $e_i(t)$, および，適正度の履歴 $D_i(t)$ を計算

$$\begin{aligned} e_i(t) &= \frac{\partial}{\partial w_i} ln(\pi(a(t), W, X(t))) \\ D_i(t) &= e_i(t) + \gamma D_i(t-1) \end{aligned}$$

5. 重み w_iの変化量$\Delta w_i(t)$ を計算

$$\Delta w_i(t) = (r_i(t) - b) D_i(t)$$

ただし b は定数．

6. 重み Wの更新

$$\begin{aligned} \Delta W(t) &= (\Delta w_0(t), \cdots, \Delta w_i(t), \cdots) \\ W &\leftarrow W + \alpha(1-\gamma)\Delta W(t) \end{aligned}$$

ただしαは非負の定数

7. $t \leftarrow t+1$ として 1 へ戻る

図 2: SGA の一般的アルゴリズム

2.3 報酬

マルチエージェントにおける報酬とシングルエージェントの報酬とを比べた場合，本質的に差があることに注意しなければならない．

シングルエージェントにおいて報酬とは，目的を達成した結果として得られる情報であるので，エージェントの存在意義として位置付けられる．

マルチエージェントの場合は報酬は個々のエージェントの目的であると同時に，さらに，マルチエージェントシステムの全体の性質を与える情報であるので，システムが効率良く働いている時には，システム全体の報酬の量が多くなるように設計する．したがって，マルチエージェントシステムは，システム全体の報酬量を最大化する問題である．各エージェントは与えられる報酬から，マルチエージェントシステム全体の報酬量を最大化するような行動を学習することが望まれる．

以上の理由から本研究ではマルチエージェントシステムにおける報酬の配分方法に関して実験，及び考察を行う．

2.4 報酬の配分方法

本研究では報酬の配分の方法を 3 種類に分類する．

case1:報酬の非共有

報酬の非共有とは図 3(a) で表されるように，それぞれのエージェントが環境から得た報酬を用いてそのまま学習する．従来より研究されているマルチエージェント強化学習では，ほとんどがこの非共有の場合である．問題点として，他のエージェントが得た報酬との相対的な報酬の量を各エージェントは知ることが出来ないために，他エージェントとの協調行動，競合回避行動等の獲得が非常に困難である点である．

case2:報酬の完全共有

報酬の完全共有とは，図 3(b) の様に各エージェントが得た報酬を全て足し合わせて，それを均質に分配し学習を行う．確率的傾斜法を用いた場合では，[木村 97] でシステム全体の割引報酬を最大化する方向に山登りすることが理論的に考察されている．完全共有では以下のような特徴が推測される．

また，完全共有の場合は [木村 97] で示されているように，マルチエージェントシステム全体を仮想的なシングルエージェントと解釈することによって，シングルエージェント時の知見を利用することができ，システム全体の割引期待報酬の最大化の方向に山登りする．

- 他エージェントの報酬情報が多い場合，自分の行動に対するノイズとなり自分の報酬獲得に関する行動の学習が遅れる．
- 各々の割引期待報酬の関数の複雑化のため，エージェント数が増加するとともに学習が遅くなる．
- 他エージェントの働きによる報酬により，さぼるような行動が獲得されるエージェントが出てくる

case3:報酬の部分共有

本研究では case1，case2 の中間である報酬の部分共有を定義し，実験，及び考察を行う．報酬の部分共有は他のエージェントの報酬を利用するが，割り引いて強化を行う．図

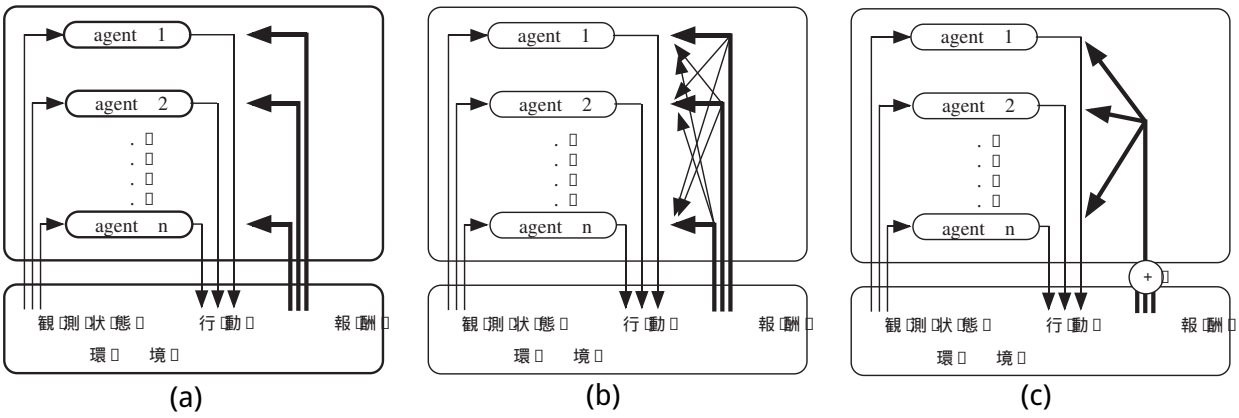


図 3: 報酬の配分方法 (a) 非共有 (b) 部分共有 (c) 完全共有

3(c) のように，各エージェントが得た報酬はそのまま強化し，他エージェントによる報酬は割り引いて強化する．協調行動の獲得と自エージェントの報酬の獲得も遅れすぎない学習法であることが期待される．

報酬の配分式

注目すべき点は，図 3(b) の他エージェントからの強化を 0.0 にすると，図 3(a) となり，逆に他エージェントからの強化を 1.0 にすると，図 3(c) となる．したがって以上をまとめ，パラメータμ $(0 \leq \mu \leq 1)$ で表すと式 (1) の様に表すことができる．$\mu = 0.0$ が報酬の非共有にあたり，$\mu = 1.0$ が報酬の完全共有で，$0.0 < \mu < 1.0$ が報酬の部分共有に相当する．

$$\begin{aligned} R_{sum}(t) &= \sum_j r_j(t) \\ r'_h(t) &= r_h(t)(1-\mu) + \mu(R_{sum}(t)) \end{aligned} \qquad (1)$$

ただし，エージェント i が環境から得た報酬を $r_i(t)$，エージェント i を強化する報酬を $r'_i(t)$ とする．

3 実験

本節では荷物の搬送を抽象化した問題を取り扱った実験を行なう．まず，実験の設定などについて説明し，計算機シミュレーションによる実験および考察を行なう．

3.1 荷物の搬送問題

本研究において荷物の搬送問題とは，次のような問題である．倉庫において搬入口から搬出口まで荷物を運ばなければならない環境を考える．タスクは，荷物を搬入口から搬出口まで運び，荷物を持っていない状態で搬入口まで戻って来ることである．複数のエージェントがこのタスクを繰り返すことによって，マルチエージェントシステムとして全体の搬出量が決定される．本実験はこの全体の搬出量を最大化することが目的である．

概念図を図 4に示す．

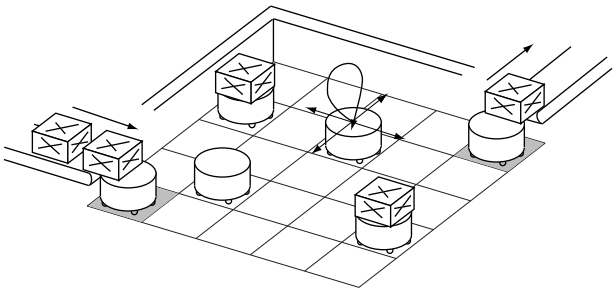


図 4: 搬出問題の概念図

実験環境は図 5(a) の様に，移動可能な環境の大きさは 5×5 のグリッドワールドとし，周囲は壁で囲まれているとする．エージェントは 1 つのグリッドに 1 エージェントで重なることは無いとし，割り振られているエージェントの番号順に行動することとする．

環境からの報酬は左後の搬入口で荷物を受け取ったときと，右前の搬出口で荷物を渡したときにそれぞれ同じ量だけ正の報酬が与えられるように設定した．往復して荷物を搬出する作業を 1 回のタスクとする．

各エージェントは感覚入力として前後左右と自分の存在するグリッドの状態を知ることができる．判別可能な種類は，壁，空白，荷物ありエージェント，荷物なしエージェントの 4 種類とする．視界が限定されていることや，他エージェントの番号を識別することができないことから不完全知覚を起こす．

行動は最大で前後左右に移動，及び停止の 5 つの行動を取ることができ，移動不可能な場所に移動する行動は選択できないこととする．

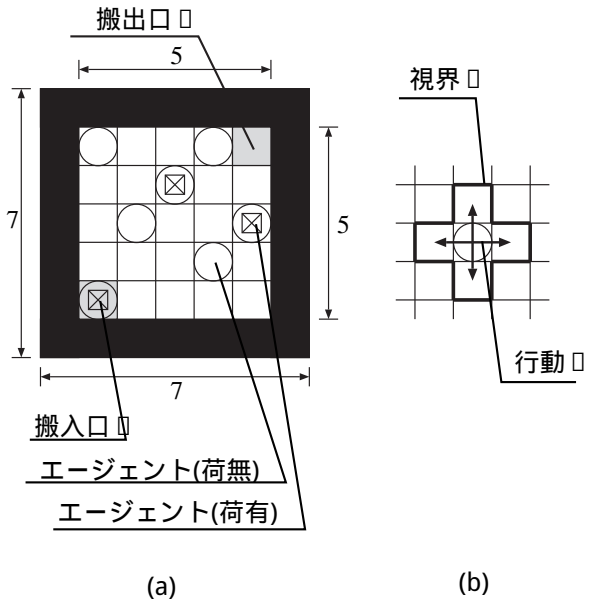


図 5: (a) 実験環境 (b) 感覚入力と行動

3.2 最適政策

この問題の最適政策は，往路と復路を別の経路として順々に移動させれば良い．エージェント数が 15 以下の場合は右廻りか左廻りかどちらかに決定し，壁に沿ってエージェント番号順に移動させれば良い．また，エージェント数が 16 以上の場合においては，15 番目のエージェント迄を，壁沿いに並べ，16 番目のエージェントからは中央の 3×3 のグリッドに配置し停止行動をとり続けさせれば良い．これによって，最適な政策時の最大の搬出量が式 (2)，(3) で計算できる．

・エージェント数が **15** 以下

$$max_task_num = 1 \div 16 \times agent_num \tag{2}$$

・エージェント数が **16** 以上

$$max_task_num = 15 \div 16 \tag{3}$$

3.3 エージェント数 5 の総搬出量

本節では，報酬の配分率 (0.0，0.5，1.0)，エージェント数 5 のときの学習ステップによる搬出量の推移について比較する．図 6は報酬の非共有，完全共有，部分共有の場合の 5000 ステップ毎の総搬出量をヒストグラムであらわしたものである．実験は 15 回行ない，その平均をグラフに表した．グラフ中の水平な破線は，節 3.2で求めた最適な場合の総搬出量である．横軸は学習ステップを表し，縦軸は 5000 ステップ毎の総搬出量を示している．

グラフから，立上りの早さと到達点において差が確認される．case1 の非共有が最も早く立上り，到達点も最適値に

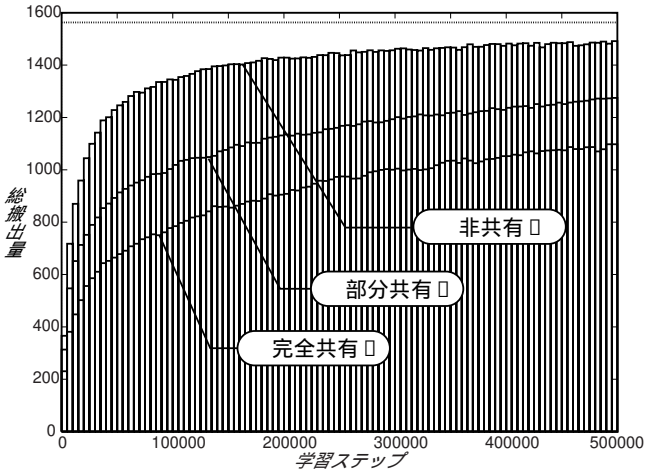


図 6: エージェント数 5 の総搬出量

近付くように学習しているのがわかる．逆に case2 の完全共有は [木村 97] によって理論的に合理性が保障されているが，立上りは遅く，500000 ステップにおいての到達点も低い．case3 の部分共有は case1，case2 の中間となった．分散は非共有が一番小さく，部分共有，完全共有と大きくなっている．

3.4 エージェント数 5 の獲得行動

報酬の共有方法の違いから獲得された，エージェント数が 5 の場合の行動の典型例を示す．

case1:報酬の非共有の場合

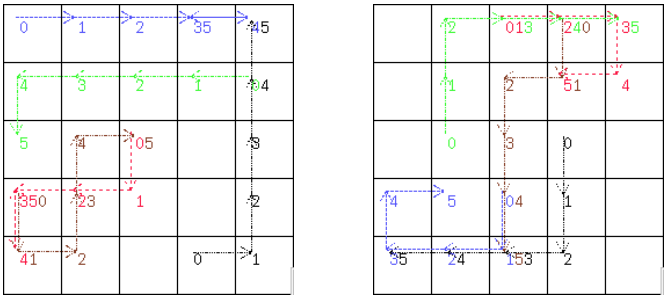
非共有の場合，図 7のような行動を獲得した．特徴としてはエージェント数が少ないため，衝突がほとんど存在しないことにより，各々報酬に向かいながら，適切な行動のうちどれかを選択することでほぼ最適な行動を学習．ただし，最適な行動を取れない状況になりうるため，最適政策ではない．

case2:報酬の完全共有の場合

完全共有の場合，図 9のような行動を獲得した．他エージェントの報酬によって学習が行われるため，働かずにさぼるような停止行動をとるエージェントが存在する．学習が遅く 500000 ステップ時においては学習が収束していないが，他のエージェントの邪魔をしない行動の学習途中であると考察でき，学習が進めば最適な行動を獲得することが期待できる．

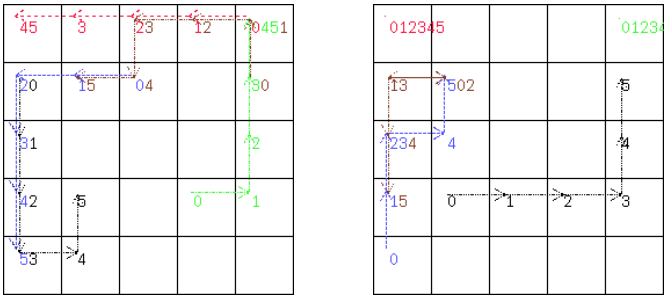
case3:報酬の部分共有の場合

部分共有の場合，図 8のような行動を獲得した．部分共有の場合でも，他エージェントの報酬により停止行動の学習がみられた．他エージェントの報酬は割り引いて考慮するため，自エージェントの報酬の獲得を優先し，立上り，500000 ステップ時における総搬出量において完全共有よりも勝っている．



(a) (b)

図 7: 非共有 ($\mu = 0.0$) の場合に得られた行動



(a) (b)

図 8: 完全共有 ($\mu = 1.0$) の場合に得られた行動

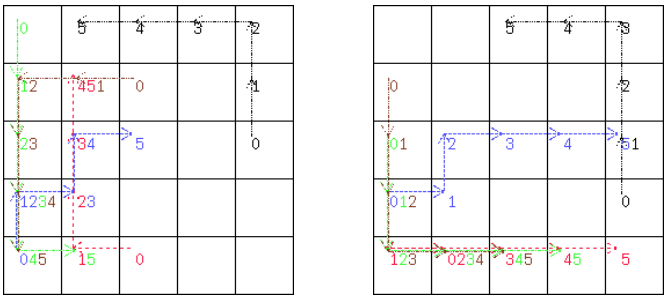
3.5 エージェント数毎の総搬出量

エージェント数を変化させた場合の総搬出量について示す．

図 10，11，12は各々非共有，部分共有，完全共有の場合の 50000 ステップ毎の総搬出量を表すグラフである．横軸はエージェント数と学習ステップを表し，50000 ステップ毎の総搬出量を表す．また，破線で表されているグラフは節 3.2で示されている最適な政策時の搬出量である．

エージェント数が 16 より少ない場合には，エージェント数が 5 の場合とほぼ同様の傾向であり，総搬出量，及び立ち上がりの早さの点で非共有，部分共有，完全共有の順である．

エージェント数が 17 以上では総搬出量で逆転が見られ，部分共有が一番高く，完全共有，非共有の順になった．ただし，エージェント数が多くなると非共有の場合分散が大



(a) (b)

図 9: 部分共有 ($\mu = 0.5$) の場合に得られた行動

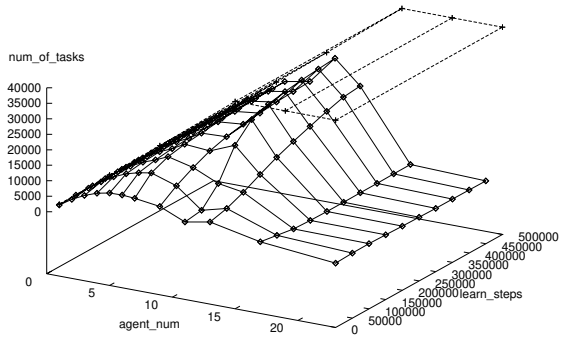


図 10: 非共有 ($\mu = 0.0$) の場合の総搬出数

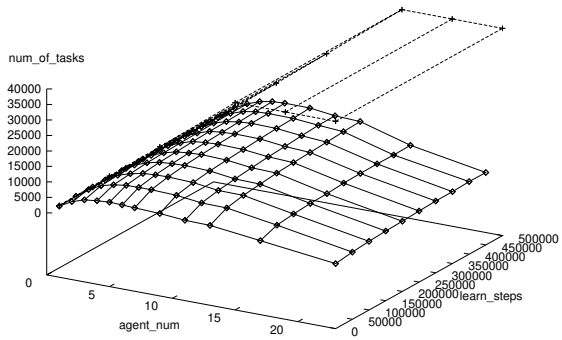


図 11: 部分共有 ($\mu = 0.5$) の場合の総搬出数

きくなる傾向が見られ，より多くの実験で比較を行なう必要がある．

3.6 考察

本研究ではマルチエージェントシステムにおいて報酬の最大化を目標として実験を行なった．エージェント数の違い等による最適行動や報酬の共有方法等について考察する．

最適行動

エージェントにとって最適な行動とは，右上に移動する場合は右か上のどちらかに移動できればよい．エージェント数が 5 の場合のように少ない場合，環境中にエージェントは少なく，衝突することがほとんどない．したがって，このような環境においては各エージェントが，自分の報酬にむかって各々勝手に学習しても，ほぼ最適な行動がとれる環境であり，このような問題クラスでは報酬の共有無しでも良い結果が得られることが実験的に示された．ただし，最適なのは節 3.2で示したように，衝突しないような経路であるため，ここで得られた行動は，最適ではない．

報酬の部分共有における利点

エージェント数が 15 以上の場合，停止行動を得る必要が

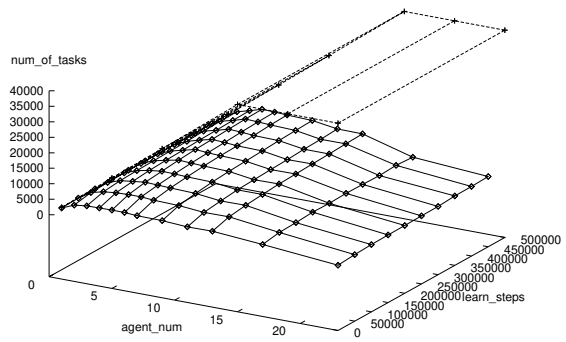


図 12: 完全共有 ($\mu = 1.0$) の場合の総搬出数

あるが，報酬の非共有の場合であれば停止行動では報酬が入らないため学習できない．報酬の完全共有であれば，他エージェントの報酬から停止行動を学習出来る可能性があるが，実際には学習が遅くなり時間がかかる．報酬の部分共有では，最適な行動を得るとは限らないが他エージェントの報酬を得ることによって，強調行動を得られ，かつ完全共有よりも立ち上がり，収束が早く実用的であると考えられる．

また，完全共有の場合，割引期待報酬の最大化の方向に山登りするが，分配を行なうだけでは局所的に安定な解に陥り，最適解に行かないことも十分に考えられる．

エージェント数大での逆転現象

今回の実験では，エージェント数 16 以上では 17，23 の 2 通りの場合について実験を行なった．エージェント数が多くなると他エージェントによる報酬と，自分の行動との相関関係の学習は非常に困難になり，実計算量も増加する．エージェント数 17 は約 10 回，23 では 3 回程度の結果をグラフに示した．エージェント数が 13 ぐらいで報酬が非共有の場合においては分散が大きくなる傾向を見せているため，より多くの実験が必要であるが，エージェント数が 23 での学習では，非共有は悪くなる傾向を見せていて，共有する場合のほうが良い傾向を見せた．この結果は，報酬を共有する必要があるケースの一例であると考えられる．

4 まとめ

4.1 本研究の成果

本研究では報酬の配分方法について，パラメータμを用いて定式化を行なった．さらに，パラメータμの値によって報酬の完全共有，部分共有，非共有に分類し，実験的考察を加えた．

本実験のような個々のエージェントが各々の報酬の最大化の方向に学習することが，マルチエージェントシステム全体の報酬を高める方向と同じ問題のクラスであれば，報酬の非共有の場合で良い結果を得ることが確認された．理論的な保障を得ている完全共有では実験的に学習が非常に遅いことが確認された．提案する部分共有では，マルチエージェントシステムに対する理論的保障は得られないが，他のエージェントとの協調動作を獲得しつつ，学習も遅くなりすぎないという点からより実用的であることを実験的に確認した．

4.2 今後の課題

今後の課題として，エージェント数と配分率の関係についての実験，及び考察が必要である．また，絶対的に協調が必要な環境についての本手法の適用を行なう必要がある．更に，配分率$\mu = 1.0$(完全共有) 以外でのモデル化と合理性についての理論的考察を行ないたい．

参考文献

[1] Watkins, C.j.C.h, and Dayan: P.Technical Note:Q-Learning, Machine Learning 8, pp.55-68(1992).

[2] H.Kimura,M.Yamamura,S.Kobayashi : Reinforcement Learning by Stochastic Hill Climbing on Discounted Reward, Proc. of 12th Int. Conf. on Machine Learning, pp.295-303(1995).

[3] 木村 元, 山村 雅幸, 小林 重信: 部分観測マルコフ決定過程下での強化学習:確率的傾斜法による接近, 人工知能学会誌,Vol.11,No.5,pp.761-768(1996).

[4] Singh, S. P.: Reinforcement Learning Algorithms for Average-Payoff Markovian Decision Processes, Proceedings of the 12th National Conference on Artificial Intelligenc, Vol.1, pp.700-705(1994).

[5] 宮崎 和光, 山村 雅幸, 小林 重信:強化学習における報酬割当の理論的考察, 人工知能学会誌,Vol.9,No.4,pp.104-111(1994).

[6] S.Arai,K.Miyazaki,S.Kobayashi:Generating Cooperative Behavior by Multi-Agent Reinforcement Learning, Proc. of 6th European Workshop on Learning Robots, pp.111-120(1997).

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

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

[9] Chrisman, L.: Reinforcement learning with perceptual aliasing: The Perceptual Distinctions Approach, Proceedings of the 10th National Conference on Artificial Intelligence, pp.183-188(1992).