目次

第 1 章　序論

1.1　ロボットの歴史

近年ロボットの研究が進むにつれて，ロボットの活躍できる分野が広がりを見せている．ロボットが人間社会で使用され始めたのは1950年代から1960年代にかけてのころである．この時代では主に工場などで使用される産業用ロボットが活躍していた [1] [2]．この時代のロボットは，人間があらかじめ動作を設定し記憶させることで，その動作を何度も繰り返し実行することができる点で注目を集めていた．この特性から，産業用ロボットが使用される場面は，単純な繰り返し作業が主であった．同じ動作を繰り返す点に関しては人間が同様の作業を行よりも効率が良いからである．しかしこの時代のロボットは単純な繰り返し動作しか行えず，状況に応じた最適な動作といった複雑な作業を行うことができなかった．

その後ロボットの制御方法に関する研究が進むことで，ロボットはより複雑な行動を実行することが可能となった [3]．その理由の 1 つとしてロボットに搭載する感覚機能が開発されたことがあげられる．この感覚機能はセンサと呼ばれる．センサは自然現象や人工物の性質やその空間情報・時間情報を，何らかの化学的原理を用いて人間や機械が扱いやすい別の媒体の信号に置き換える装置のことである．ロボットにセンサが搭載されることで，ロボットは自身が置かれている現在の状態を数値情報として認識することができる．その数値情報に応じて自身の実行する行動を選択，変化させることが可能となった．またロボットが自身の行動を制御可能となったことで，ロボットが使用される場面もさらに増大した．

現在では，何らかの方法で得た知識を利用することで学習を行い，自身の実行する行動に反映させる学習制御ロボットの研究が行われている．学習制御ロボットの研究が進めば，事前に状況に合わせた最適な動作を設定しなくとも，ロボット自身が様々な情報を得ることで自身の行動制御方法を自律的に獲得することが可能となる．学習制御ロボットが実用化される段階になれば，ロボットが活躍することができる場面もさらに拡大することが期待される．このロボットに学習制御機能を持たせる研究はロボット工学における機械学習の研究分野に属される．本研究はこの機械学習に関する研究を取り扱っている．

1.2　学習による自律的行動獲得

機械学習とは，人間が自然に行っているパターン認識や経験則を導き出したりする活動を，コンピュータを使って実現するための技術や理論，またはソフトウェアの総称である．機械学習には様々な手法が存在する．大きく分けて教師あり学習，教師なし学習，強化学習の 3 種類に分けられる．

教師あり学習とはクラスラベルや閾値などの学習すべき付随情報がデータと共に事前に人間から与えられる．付随情報がないデータが与えられた時に，対応する付随情報を予測するための関数や規則を獲得する手法が教師あり学習である．

教師なし学習では事前にデータと関係のある学習すべき付随情報は与えられない．与え

られたデータの分布などからデータの特徴的なパターンを見つける学習手法である．具体的な手法としてクラスタリングやパターンマイニングなどが存在する．

強化学習とはある環境における最適な行動を，実際に経験することで学習するタイプのアルゴリズムである [4] [5] [6]．強化学習の概要を図 1 に示す．強化学習では学習を行う存在をエージェントと呼ぶ．エージェントは行動すると環境から報酬と呼ばれるスカラ値を受け取る．この報酬の累計をできるだけ多くする行動戦略を獲得することを目的とした学習方法が強化学習である．強化学習は自律エージェントや自律ロボットの学習制御アルゴリズムとして注目されている [7]．その理由の 1 つとして上げられるのは強化学習が実環境における最適行動を獲得することに優れている点である．自律エージェントや自律ロボットは実世界の変化する環境で動作する．そのため，環境との相互作用を通して学習する強化学習の枠組みが適している．本研究では機械学習の中でも，実ロボットへの適用性が高い強化学習を中心に扱う．

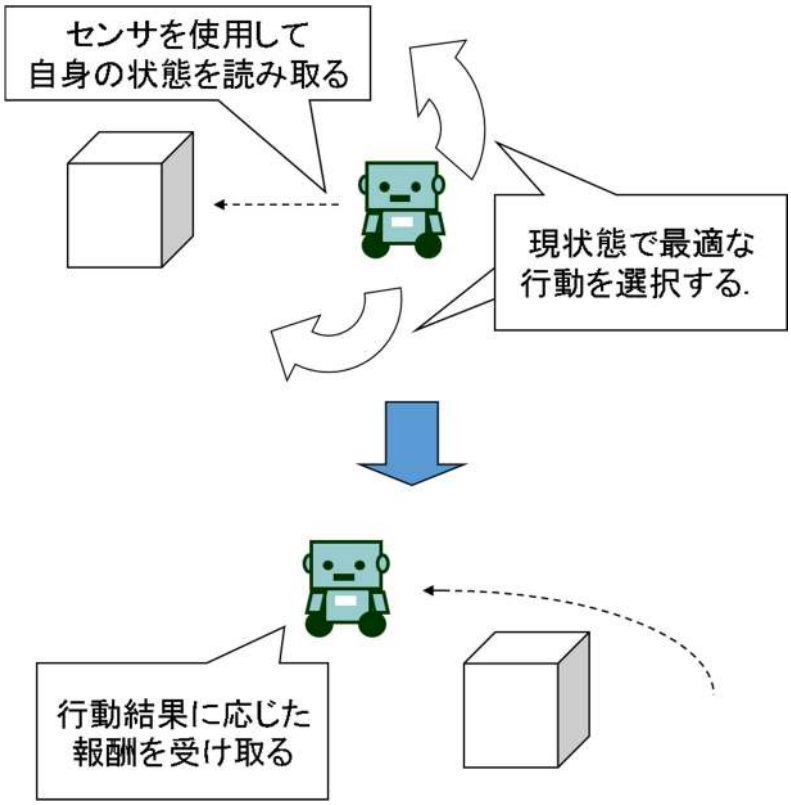


図 1：強化学習の概要

1.3 マルチエージェントシステム

マルチエージェントシステムとは，多数のエージェントによって構成されるシステムである [8] [9]．マルチエージェントの概要を図 2 に示す．それぞれのエージェントは自身の置かれている環境状態を知覚して，自身に与えられた目的を達成するように行動する．マル

チエージェントシステムの大きな特徴の 1 つとして，システム全体の振る舞いはエージェント同士が相互に作用することによって決定される点が上げられる．またシステム全体の振る舞いは各エージェントの行動決定に影響を及ぼす．そのためマルチエージェントシステムでは各エージェントが他のエージェントの状態を認識し，他のエージェントに合わせた行動選択を行うことが重要となる．

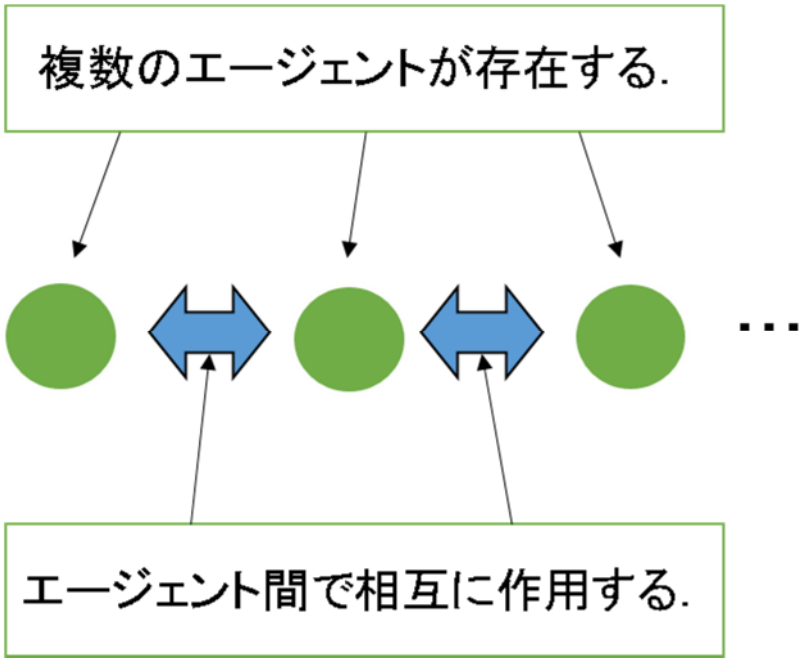


図 2：マルチエージェントシステムの概要

マルチエージェントシステムの各エージェントに対して機械学習による自律的行動獲得に適用する場合もある [10] [11] [12]．マルチエージェントシステムを用いた機械学習とは，機械学習を行うエージェントを複数用いたシステムである．マルチエージェントシステムを用いた機械学習を適用することで，環境に適した行動だけでなく複数のエージェントが存在する環境内で各エージェントが他のエージェントとの相互作用を学習することができる [13]．またエージェント間で経験情報や知識等を共有することで学習に必要とされる時間を短縮するという利用方法も存在する [14]．本研究ではマルチエージェントシステムを機械学習に適用することで，エージェントが複数存在する環境内で各エージェントとの相互作用を学習する手法について扱う．

1.4　先行研究

我々はこれまでにマルチエージェントシステムによるシングルロボットの行動学習手法を提案した [15]．マルチエージェントシステムによるシングルロボットの行動学習手法の概要を図 3 に示す．この先行研究ではロボットの行動をロボットに搭載されているアクチ

ュエータの動作の組み合わせによって決定されるものとする．したがってロボットの行動はアクチュエータの動作に分けることができる．そこで先行研究では単体ロボットに複数のエージェントを設定し，各エージェントにそれぞれロボットに搭載されているアクチュエータの動作を割り当てる．各エージェントは自身に割り当てられたアクチュエータの動作を学習する．これによって１エージェントでロボットの行動を学習する場合と比較して，１つのエージェントに割り当てられた行動数が削減される．その結果１エージェントでロボットの行動を学習する場合よりも，最適行動を学習するまでにかかる時間を短縮する手法である．

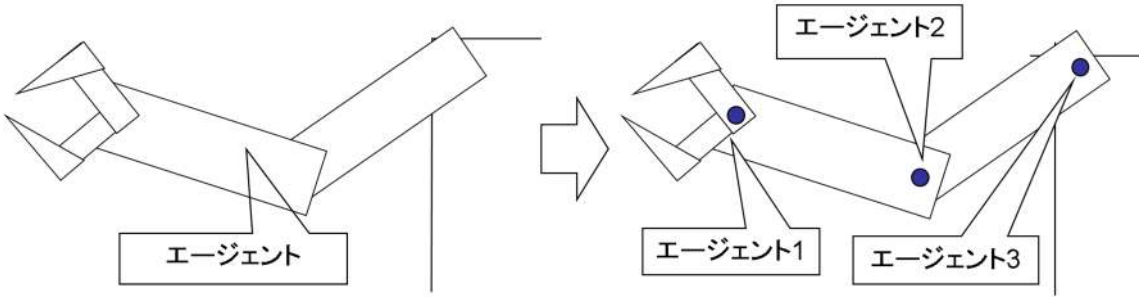


図 3：マルチエージェントシステムによるシングルロボットの行動学習手法の概要

マルチエージェントシステムによるシングルロボットの行動学習手法では，各エージェントはロボットに搭載されたアクチュエータの動作を学習する．そのためロボットに搭載されるエージェントの数は最大でロボットに搭載されているアクチュエータの数となる．マルチエージェントシステムによるシングルロボットの行動学習手法の処理の流れを図 4 に示す．ロボットは現在の環境状態を取得後，各エージェントに環境状態の情報を送信する．各エージェントは現在の環境状態を元に自身に割り当てられたアクチュエータの動作を選択する．全エージェントが行動を選択後，各エージェントが選択した行動をロボットの行動として出力する．ロボットが行動した後，ロボットは環境から報酬を受け取る．受け取った報酬は各エージェントに送信され，各エージェントは報酬を元に自身が選択した行動を学習する．この一連の流れを繰り返すことで各エージェントは自身に割り当てられたアクチュエータの動作を学習する．

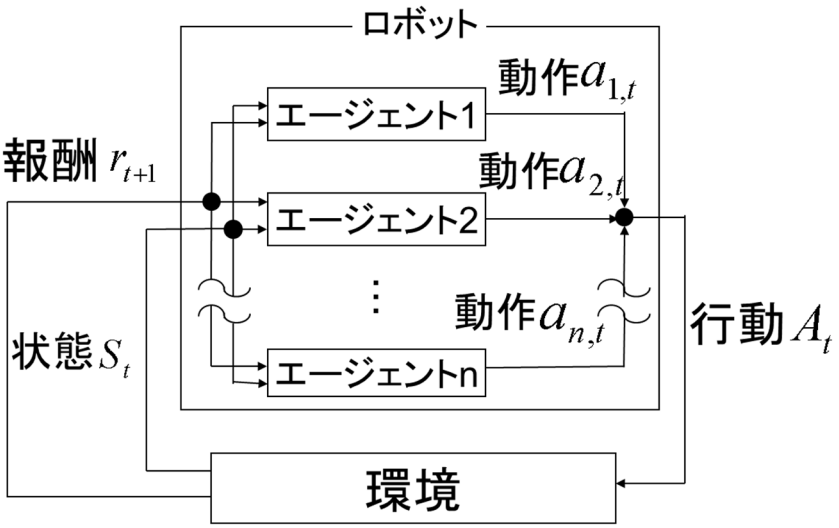


図 4：マルチエージェントシステムによるシングルロボットの行動学習手法の処理の流れ

1.5　先行研究の問題点

先行研究の問題点の 1 つとして，各エージェントが他のエージェントとの協調行動を行うメカニズムが無い点が上げられる．各エージェントは現在の環境状態から行動を選択する．そのためロボットが環境情報のみでは行動を一意に決定できないような状態である時，各エージェントの行動選択によってその状況に応じた最適行動を選択できない可能性が発生する．

各エージェントが他のエージェントとの協調行動を行うメカニズムが無いことによって発生する障害の例として，現在の環境状態におけるロボットの最適行動が複数存在する場合を上げる．ロボットの最適行動が複数存在する場合での先行研究の問題点によって発生する事象の例を図 5 に示す．図 5 では 1 つのタスク達成状態に対して最適行動が 2 つ存在する場合の例を示している．この例ではロボットが置かれている環境状態では 2 つの最適行動が存在するため，どちらかの行動を出力するべきである．しかし各エージェントから見た場合 2 つの最適行動の内 1 つに決定するための情報が存在しない．そのため各エージェントの意思決定が揃わないため，各エージェントの選択した行動によって最終的にロボットが出力する行動が最適行動とは異なるという状況が発生する．特にロボットの最適行動の数やエージェントの数が増えるにつれて，エージェント間の意思決定が揃わなくなる確率は高くなる．したがってロボットが最適行動を出力できる確率も下がるという状況になる．

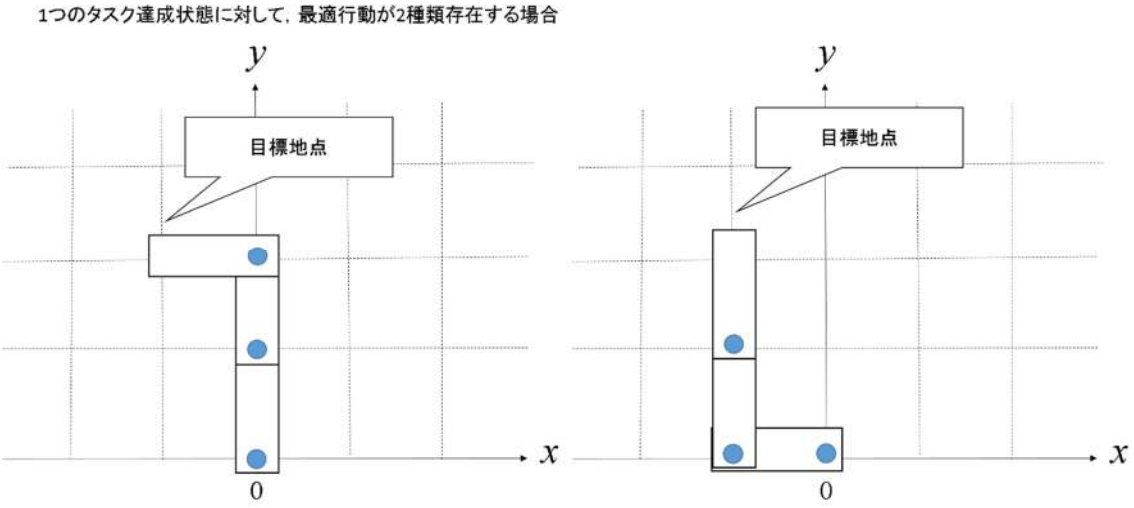


図 5：ロボットの最適行動が複数存在する場合での
先行研究の問題点によって発生する事象

この問題点の原因として各エージェントが意思疎通を行っていない点があげられる．マルチエージェントシステムの特徴の 1 つとしてシステム全体の振る舞いがエージェント同士の相互作用によって決定される点がある．先行研究の場合ではシングルロボットの行動は各エージェントが選択したアクチュエータの動作によって決定されることになる．そのため各エージェントが行動を決定するためには環境状態だけでなく他のエージェントの情報も考慮して決定しなければならない．そのためロボットの最適行動が複数存在するタスクを実行する際には,各エージェントが協調することで複数存在する最適行動の中から 1 つを選べるようにする必要がある．

1.6　従来研究とその問題点

マルチエージェント強化学習によるロボットの行動獲得手法としては,「知覚情報の素視化によるマルチエージェント強化学習の高速化-ハンターゲームを例に- [16]」「災害救助を目的とした低電力消費型ロボットネットワークシステム [17]」「複数台移動ロボットによる協調獲得 [18]」等が存在する．これらの研究は 1 ロボット 1 エージェントを対象としたマルチエージェント強化学習に関する研究である．これらのようにマルチエージェントシステムをロボットに適用した研究は，1 ロボット 1 エージェントを対象にした研究が多い．1 ロボット 1 エージェントを対象としたマルチエージェントシステムと 1 ロボット多エージェントを対象としたマルチエージェントシステムではエージェント間の協調行動や通信で交換する情報等が異なる．その理由としては本研究とはエージェントの単位が異なる点である．エージェント単位が異なると各エージェントが必要とする情報や選択した動作が他のエージェントに与える影響が異なる．そのため 1 ロボット 1 エージェントを対象としたマルチエージェントシステムに関する手法をそのまま 1 ロボット多エージェントを対象と

したマルチエージェントシステムに適用したとしても十分なパフォーマンスを発揮することができない.

1 ロボット多エージェントを対象としたマルチエージェントに関する研究としては「6 軸マニピュレータの分散制御実験 [19]」「多自由度機構と分散制御 [20]」「局所的な齟齬情報に基づくヘビ型ロボットの適応的自律分散制御方策 [21]」等が存在する．これらの研究では単体ロボットの行動獲得手法としてマルチエージェントシステムが使用されている．しかしエージェントの行動戦略に関しては設計段階で決められており，エージェントが自律的に行動獲得はしない．そのため 1 ロボット多エージェントを対象としたマルチエージェント強化学習による自律的行動獲得に関する手法は研究されていないという状態である．

このように単体ロボットの各アクチュエータの協調行動を機械学習によって自律的に獲得することを目的としたマルチエージェントシステムに関する研究は，現在進んでいないというのが現状である．しかし 1 ロボットに搭載されるアクチュエータの数やセンサの数が増加すれば，人間による行動設計や 1 エージェントによる行動学習はますます困難となってくる．そのためマルチエージェントシステムを用いて単体ロボットの行動を学習する事は非常に重要となってくる.

1.7 本研究の目的

本研究では単体のロボットにマルチエージェントシステムを適用し，各エージェントが他のエージェントと協調した行動選択を学習するシステムを提案する．この手法を単体のロボットに適用することで，ロボットの置かれている環境状態だけでは行動を一意に決定できない状況の際に，各エージェントが協調することで 1 つの行動を選択することができることを目指す.

1.8 問題解決のアプローチ

問題点を解決する方法の 1 つとして，各エージェントが他のエージェントとの協調行動を学習することでエージェント間の意思疎通を行う方法が上げられる．そのためには各エージェントが他のエージェントと協調する方法を考える必要がある.

エージェント間の協調動作の学習方法の 1 つとして，各エージェントの取得する状態にロボットが取得する環境状態に加えて他のエージェントが選択した行動を加える方法が考えられる．この方法を使用することで各エージェントは他のエージェントが選択した行動を状態の 1 つとして認識し，行動選択や学習の際にロボットの取得した環境状態と他のエージェントが選択した行動に対する自身の行動を選択，学習することができる．しかしこの方法の問題点として行動選択時に，各エージェントが同時に行動選択を行うと他のエージェントが選択する行動を知ることができないという点が存在する．また各エージェントが行動選択を行うタイミングを独立に行った場合では，先に行動選択を行うエージェントはまだ行動選択を行っていないエージェントが選択する行動を知ることができない．そのた

め各エージェントが選択する行動を情報として利用することができなくなる．

そこで本研究ではマルチエージェントシステムで各エージェントが他のエージェントの選択した行動を状態として利用する方法の１つとして，ロボットの１回の行動選択の際に，ロボット内で各エージェントの行動選択を何度も仮想的に行うことで各エージェントの行動を決定する方法を考える．この複数回行動選択を行い各エージェントが選択した行動を情報として交換する一連の動作を反復合議に基づく協調アルゴリズムと呼ぶ．反復合議に基づく協調アルゴリズムのアプローチを図 6 に示す．この方法を適用することで各エージェントが行動選択を行い，選択した行動を情報として他のエージェントに送信することができる．したがって各エージェントは行動選択の際に他のエージェントが選択した行動情報を状態として利用することができるようになる．行動選択時の各エージェントの行動選択と行動情報の送信に関する説明は第 3 章で詳しく説明する．

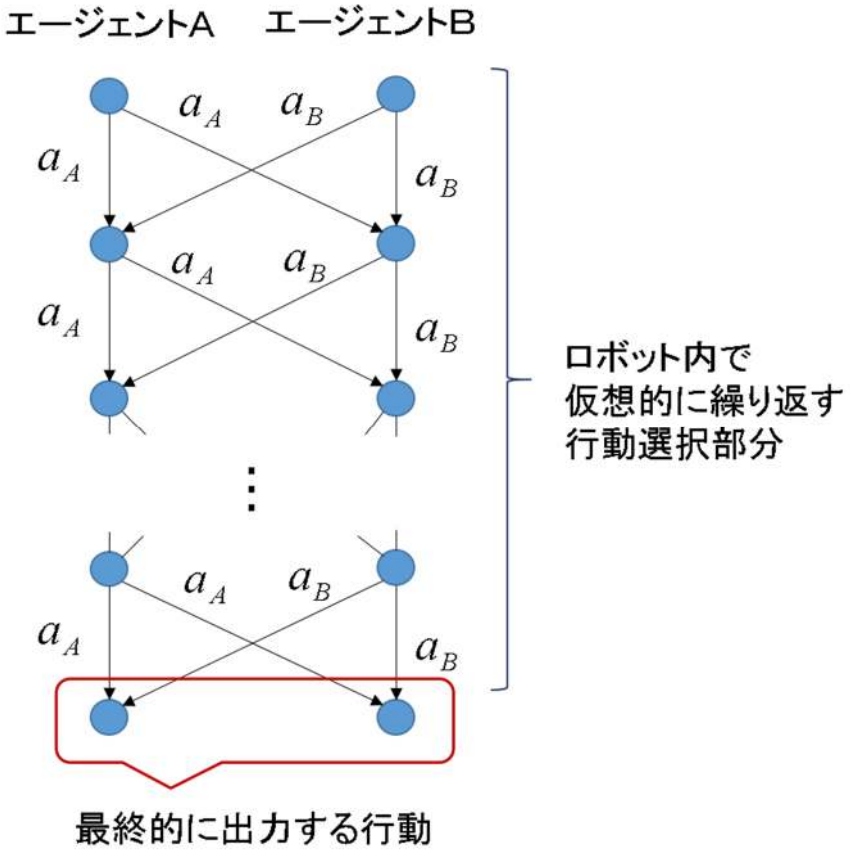


図 6：反復合議に基づく協調アルゴリズムのアプローチ

1.9　本論文の構成

第 1 章ではロボットの制御方法の背景から現在研究されている機械学習と強化学習，またマルチエージェントシステムについて説明した．また我々が行った先行研究の 1 つであるマルチエージェントシステムによるシングルロボットの行動学習手法について説明し，

その問題点について述べた．そして先行研究の問題点から本研究の目的を述べ，目的達成のアプローチを示した．

第 2 章ではマルチエージェントシステムに関する基本的な枠組みと，本研究で使用するマルチエージェントシステムについて述べる．またマルチエージェントシステムにおけるエージェント間の協調について説明する．

第 3 章では本研究で提案する反復合議に基づく協調アルゴリズムについて説明する．

第 4 章では本研究で行う実験について説明する．まず初めに実験を行う目的を述べ，実験の概要について説明する．そして実験結果を示し，その実験結果から考察を述べる．

第 5 章では論文全体のまとめを述べる．また提案手法の今後の課題について述べる．

第 2 章　マルチエージェントシステム

本章では，マルチエージェントシステムについて説明する．始めに一般的なマルチエージェントシステムの定義について述べる．次に本研究で使用するマルチエージェントシステムについて説明する．最後にマルチエージェントシステムにおけるエージェント間の協調について説明する．

2.1　マルチエージェントシステムの概要

マルチエージェントシステムとは，複数のエージェントが集まり，共通の目的をもって，作業や処理などを行うシステムである [8] [9]．マルチエージェントシステムの基本的な構成を図 7 に示す．マルチエージェントシステムの最大の特徴は，複数のエージェントから構成される分散システムということである．ただし，何がどのように分散しているかに関しては，マルチエージェントシステムを適用する手法やシステムなどに応じて異なる．また各エージェントが出力する行動や処理に使用する情報や知識等も分散されている場合がある．

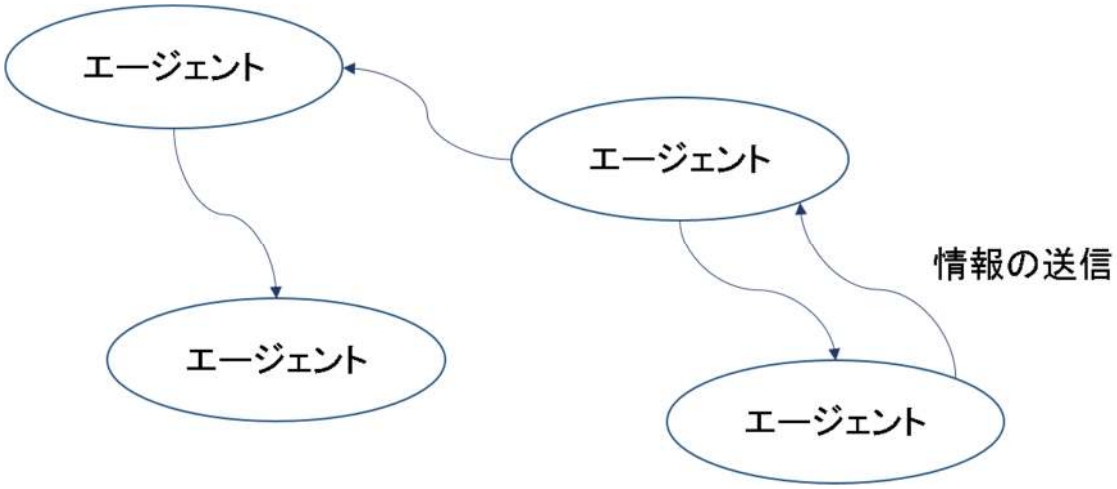


図 7：マルチエージェントシステムの基本的な構成

マルチエージェントシステムのもう 1 つの大きな特徴は，各エージェントがばらばらに動作するのではなく，エージェント全体が共通の目的に向かって，整合性のある相互作用的行動を行うところである．エージェントが相互作用的行動を行う際のエージェント間の関係には，互いに通信を行ったり，統合したりする形態が考えられる．このようなエージェント間の関係から整合性のある全体の挙動を得ることが，マルチエージェントシステム特有の技術である．

マルチエージェントシステムには，柔軟性，頑健性，効率的処理などの利点が期待されている．一方で整合性，最適性，動機などの別の問題も発生する．そのためシステムをマルチ

エージェント化させるかどうかは，条件を十分に考慮して判断すべきである．

マルチエージェントシステムに関する研究は非常に多岐に渡っており，それぞれ異なる技術領域，学問分野を背景としている．そのため研究分野に応じて様々なシステムに応用されている．マルチエージェントシステムを用いたロボットシステムに関するものとしては，群ロボットの研究 [22]，協調制御の研究 [23]，分散制御の研究 [19]やモジュール化ロボット [18]の研究が存在する．これらの研究ではエージェントの役割から出力する行動，共有する情報が異なる．マルチエージェントシステムを使用する際には，エージェントの役割や情報共有について明らかにすることが重要である．

2.2 本研究で使用するマルチエージェントシステム

本研究で使用するマルチエージェントシステムでは，単体のロボットに対して使用するシステムとする．各エージェントが出力する行動は，ロボットに搭載されているアクチュエータの動作とする．各エージェントにはロボットに搭載されているアクチュエータ 1 つまたは複数の動作を割り当てる．各エージェントが取得する環境状態に関する情報は共通のものとする．また各エージェントが所持する状態行動対は各エージェントが独立で所持する．

本研究で使用するマルチエージェントシステムでは，各エージェントは自身に割り当てられたアクチュエータの動作を学習する．各エージェントは現在の環境状態から自身に割り当てられた行動の中から最適と考える行動を選択する．エージェント 1 つだけではロボットに搭載されたアクチュエータの 1 部の動作しか出力できない．だがロボットに搭載されている全アクチュエータの動作をロボットに搭載するエージェントに分配することで，全エージェントが行動を設定すれば全アクチュエータの動作を出力することができるようになる．ロボットに搭載されている全アクチュエータを動作させることが可能であれば，ロボットが実行可能である全行動を出力することができる．したがってロボットに搭載されているアクチュエータの動作を複数のエージェントの出力という形で分配した状態でも，ロボットの出力可能な全行動を出力可能な状態になるということである．

2.3 エージェント間の協調的行動

マルチエージェントシステムの大きな特徴の 1 つとして，エージェント間で協調的行動を行うことで，システム全体で 1 つの目標を達成する点が存在する．エージェント間で協調的行動を行うためには，エージェント間で何らかの情報交換や情報共有を行わなければ，各エージェントは独立で目標に向かって行動選択を行う．各エージェントが独立で行動選択を行うと他のエージェントの行動の妨害となる行動を選択する場合がある．しかし各エージェントは他のエージェントとの協調的行動を行わないため他のエージェントの妨害となる行動を改めることができない．その結果システム全体で目標を達成することができなくなるのである．特に本研究では単体のロボットに対してマルチエージェントシステムを

適用する．この場合では各エージェントが単体のロボットという形で物理的に連結している．そのため各エージェントが選択した行動は他のエージェントの状態に影響を与える可能性が高い．したがってエージェント間の協調的行動は非常に重要になってくる．

本研究で使用するマルチエージェントでは，行動選択のタイミング，行動出力のタイミング，エージェントが選択した行動情報の共有という方法でエージェント間の協調的行動を実現する．行動選択のタイミングは，全エージェントが行動選択を行うタイミングを，同期をとることで全エージェントが同時に行うということである．これにより全エージェントが取得する環境状態が同じ時間帯のものとなる．そのため各エージェントが行動選択を行う際に他のエージェントと協調的行動を行える状態となる．行動出力のタイミングは，各エージェントが行動を選択後すぐに出力するのではなく，全エージェントが行動選択を行い出力する行動が決定した後，全アクチュエータが同時に出力する．これによって各エージェントが行動選択時に行動選択前と行動選択後に行動を出力するタイミングで環境状態異なるという状況を防ぐ．エージェントが選択した行動情報の共有とは，各エージェントが選択した行動をほかのエージェントに環境状態という扱いで送信する．これを全エージェントが行うことでエージェント間の行動情報の共有を行う．これによって各エージェントは他のエージェントが選択した行動を環境情報として認識し，行動選択を行うことができる．エージェント間の行動情報の共有については第 3 章で詳しく説明する．

第 3 章　反復合議に基づく協調アルゴリズム

本章では，本研究で提案する手法に関する内容を説明する．初めに本研究で提案する手法の概要を説明する．そして提案手法の詳しい構成，詳細について説明する．また本研究で使用した行動選択方法について説明する．

3.1　提案手法の概要

本研究では，単体ロボットの意思決定手法としてマルチエージェントシステムを用いた強化学習手法を提案する．提案手法の概要を図 8 に示す．本研究では単体ロボットに複数のエージェントを設定する．各エージェントにはロボットに搭載されているアクチュエータ単位の動作を割り当てる．各エージェントはロボットが認識する環境状態 S に加えて，他のエージェントが選択した行動 a^i を状態として認識し行動選択を行う．i は各エージェントの番号とする．この行動選択をロボットに搭載されている全エージェントで行いロボットが実行する行動を決定する．行動学習時には各エージェントが環境状態 S と他のエージェントが選択した行動 a^i の際に自エージェントが行動 a を選択したときにどれだけの報酬 r を得られたかを学習する．このように各エージェントが環境状態 S に加え，他のエージェントが選択した行動 a^i を状態として認識し，行動選択や学習を行うことで他のエージェントの選択する行動と協調する行動を選択することができるようになる．

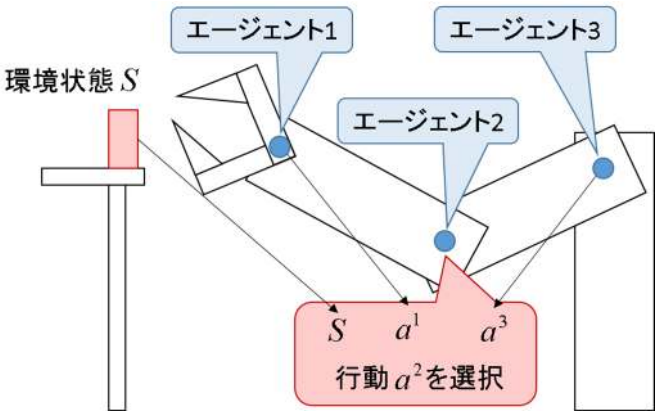


図 8：提案手法の概要

本研究で提案する手法ではロボットが実行する行動選択 1 回に対して各エージェントの行動選択を仮想的に複数回行う，この各エージェントが行う行動選択をステップ sp と定義し，ロボットの行動を決定する際にはステップを何回も繰り返すものとする．1 回の出力する行動を決定する際に，各エージェントが複数回行動選択を行う理由は，各エージェントの選択した行動の情報を交換する必要があるためである．本研究で提案する手法は各エージ

ェントの行動選択の際に環境状態 S に加えて他のエージェントが選択した行動 a^i を状態として認識する．しかし各エージェントの選択した行動 a^i を認識するためには，各エージェントが行動選択を行わなければならない．そこでロボットが出力する行動を決定する際に各エージェントが複数回行動選択を行い，行動選択後に各エージェントが選択した行動 a^i の情報を交換し共有する．行動選択と情報共有を繰り返すことで各エージェントが他のエージェントと協調した行動選択を行い，既定のステップ回数を満たした時にロボットの最適な行動を決定する．本論文中ではエージェント C の B 番目の行動を ${a_B}^C$ と表記する．

提案手法における 1 回の出力行動決定の概要を説明する．1 回の出力行動決定時における各エージェントの行動選択と他のエージェントの協調について図 9 に示す．ロボット内には m 個のエージェントを設定する．ロボットは現在の環境状態 S を取得する．始めにステップ 0 では各エージェントはロボットが取得した環境状態を元に各エージェントに割り当てられた行動 a^i を選択する．全エージェントが行動 a^i を選択後，ステップ数を 1 つ増やす．ステップ 1 以降からは各エージェントに他のエージェントが前ステップで選択した行動 a^i の情報を送信する．各エージェントが他のエージェントの行動情報を取得後，各エージェントは現在の環境状態 S と他のエージェントが前ステップで選択した行動 a^i を状態として，行動 a^i を選択する．各エージェントが行動 a^i を選択後，ステップ数を 1 つ増やす．以降ステップ数が任意の規定回数 N 回に到達するまでステップ 1 と同様に情報交換と行動選択を行う．ステップ数が N 回に到達したとき，各エージェントが選択した行動 a^i が，実際にロボットが出力する行動 $A=\{a^1,a^2,\cdots,a^m\}$ となる．

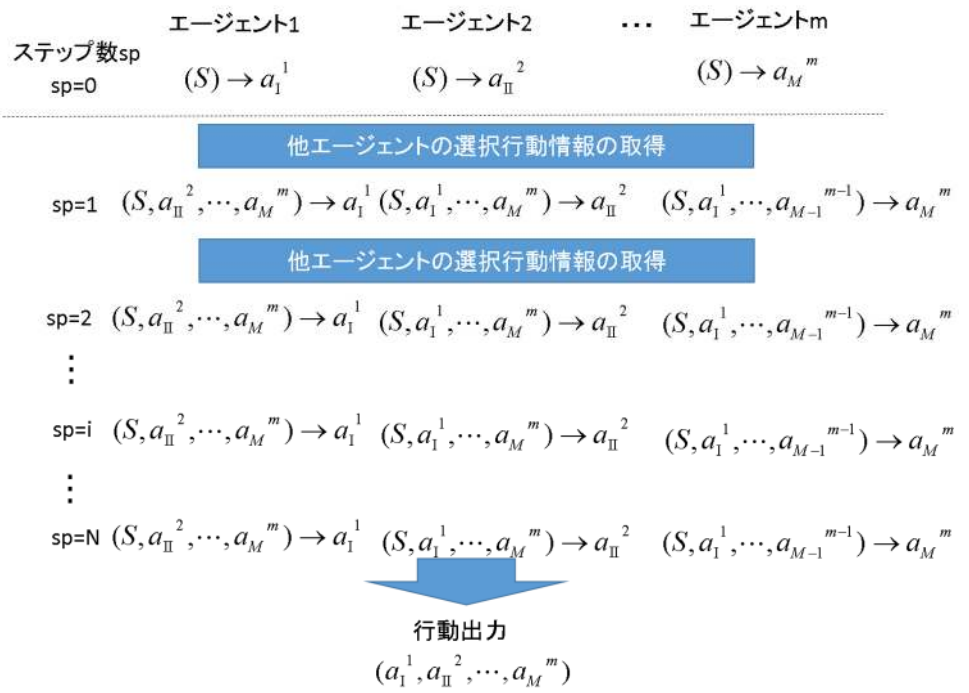


図 9：提案手法の各エージェントの行動選択と協調

3.2 提案手法の構成

本研究で提案する手法を適用したロボットの行動選択と行動学習の流れについて説明する．提案手法を用いたロボットの行動選択から行動出力の流れを図 10 示す．本研究で提案する手法は，単体ロボットの行動獲得の際に適用できる手法である．ロボットは環境状態 S を認識した後，環境状態 S を各エージェントに送信する．各エージェントは環境状態 S を元に任意の行動選択手法に基づいて行動を選択する．各エージェントが行動選択手法に基づいて選択した行動 a^i は他のエージェントに情報として送られる．その後各エージェントは環境状態 S と他のエージェントが選択した行動 a^i を元に行動選択を行う．この過程をステップ数が任意のステップ回数 N に到達するまで行い，ステップ数が N 回に到達したとき，各エージェントが選択した行動 a^i が，ロボットが実際に出力する行動 A となる．ロボットが行動 A を出力後，環境からタスクの達成度に応じた報酬 r を受け取る．各エージェントは報酬 r を受け取り，環境状態 S において他のエージェントが行動 a^i を選択した際に，自身が選択した行動 a^i について学習する．この一連の流れをロボット 1 回の行動選択と学習とし，ロボットに与えられたタスクを達成するまで繰り返す．

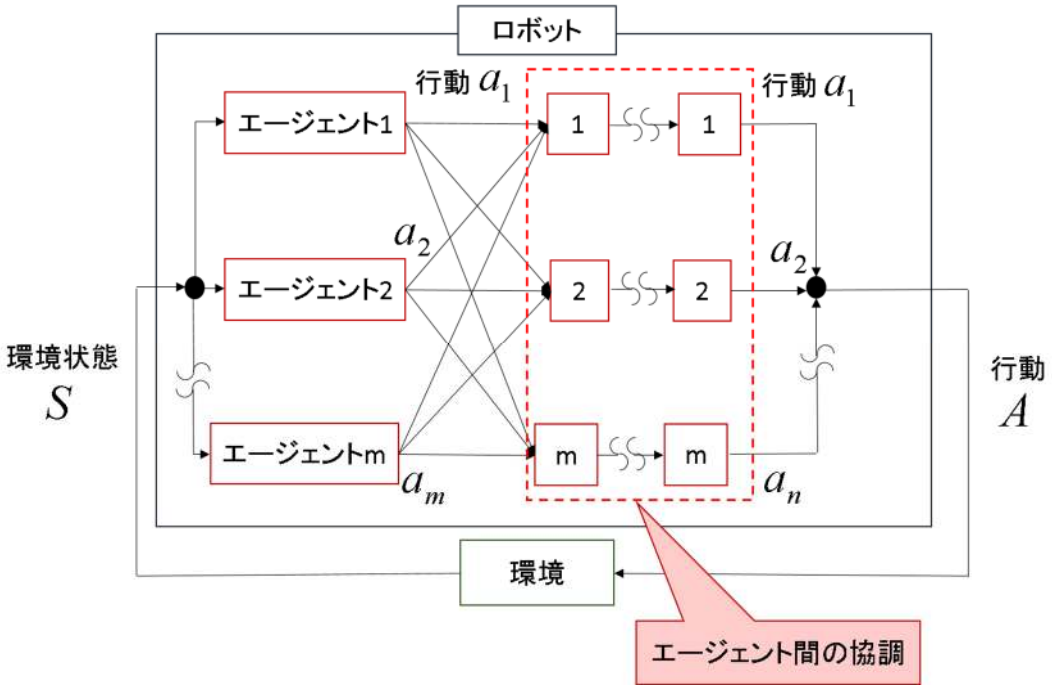


図 10：提案手法を用いたロボットによる行動選択の流れ

各エージェントはそれぞれ状態行動対に対して任意の学習法に基づいた行動評価値を所持する．本研究では強化学習手法の 1 つである $Q-learning$ 法を用いる．そのため各エージェントはそれぞれ状態行動対に対して Q 値を所持する．状態行動対の内，状態 S は環境状態と他のエージェントが選択した行動 a^i となる．行動 a^i は各エージェントに割り当てられたアクチュエータの動作となる．

各エージェントは任意の行動学習方法によって求められた行動評価値を元に行動選択を行う．行動選択の際には常に最適と思われる行動を選択するのではなく，ある一定の確率で探査的行動を行う必要がある．その理由としてはエージェントが未経験の状態や行動の中に現在までに学習した状態行動対より適した行動が存在する可能性があるためである．最適行動選択を行うか探査的行動を行うかを決める際にどの範囲で決定するかは表 1 に挙げている範囲が考えられる．本研究ではエージェント単位で最適行動か探査的行動を行うか決定する．エージェント単位の探査行動について図 11 示す．各エージェントが環境状態を取得後，最適行動を選択するか探査的行動を選択するかを決定する．以後出力行動選択 1 回の間では探査的行動が選択されたエージェントは始めに選択した行動 a^i を選択し続ける．同じ行動を選択し続ける理由は，探査的行動を選択するエージェントがステップ毎にランダムに行動選択を行うと，他のエージェントが毎ステップでランダム行動に合わせた行動を選択しなければならないため 1 つの行動に決定できないからである．

表 1：行動選択時の探査的行動選択の範囲

ロボット単位	ロボットの 1 回の行動選択時に選択
エージェント単位	各エージェントそれぞれがロボット 1 回の行動選択時に選択
各ステップの各ステップの行動選択単位	各エージェントの各ステップの行動選択時に選択

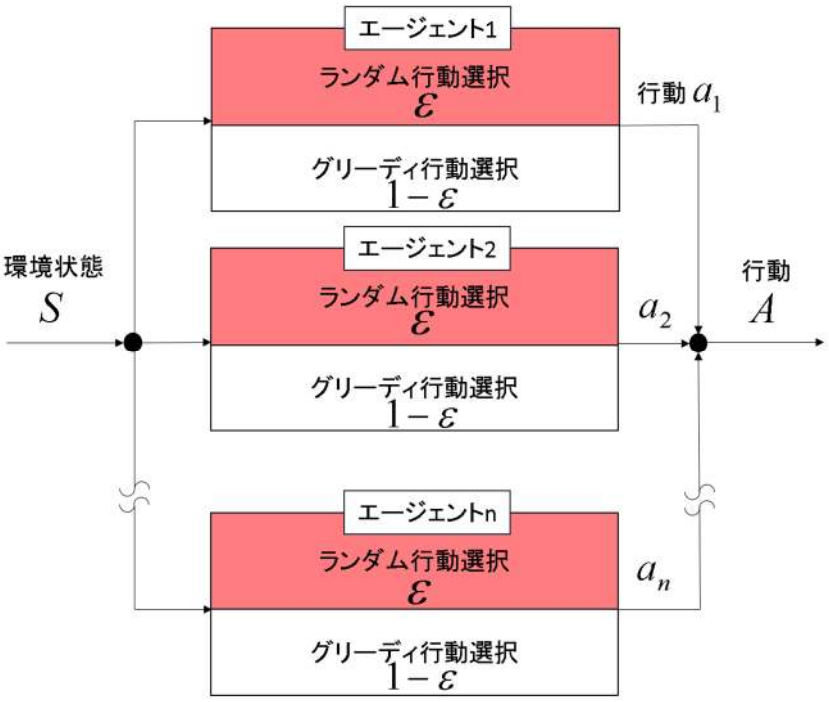


図 11：エージェント単位による探査的行動選択

3.3 行動遷移確率と協調的行動評価値に基づいた行動選択方法

本研究では各エージェントの各ステップ sp における行動選択の際に行動評価値 Q 値に加えて別の行動評価指標を用いて行動選択を行う．行動評価値のみで行動選択を行うと最適な行動を学習済みであるにもかかわらず最適行動を選択できないという状況が発生する．その理由として行動評価値のみで行動選択を行うと，他のエージェントの選択した行動の情報は前ステップ 1 回で選択した行動の情報しか利用できないためである．直前のステップで選択した行動の情報しか行動選択時に考慮できないと各エージェントが常に他のエージェントが直前に選択した行動に合わせた行動選択を行う．この場合では常に各エージェントの意思決定が揃わないまま既定ステップ数 N を満たし行動選択が完了してしまう可能性がある．この状態が発生したときに選択した行動は学習初期や 0 ステップ目での行動選択

時の乱数に左右されてしまう．そこで各エージェントが選択した行動情報から各エージェントが選択する行動の確率を求めて利用する手法を提案する．この手法によって各エージェントの行動選択が乱数に左右されず，他のエージェントが選択する確率の高い行動に合わせた行動選択が可能となる．

本研究では各エージェントの行動選択の確率を行動遷移確率 $\pi_{a^j}(a^k)$ と定義する． a^j はエージェント j が選択した行動， a^k はエージェント k が選択した行動である．行動遷移確率 $\pi_{a^j}(a^k)$ は各エージェントの Q 値を元にロボットの 1 回の行動選択開始時に各エージェントに対して計算する．行動遷移確率 $\pi_{a^j}(a^k)$ の計算方法を図 12 に示す．エージェント j の行動遷移確率 $\pi_{a^j}(a^k)$ は現在の環境状態 S で各行動 a^j を出力した際の最大の Q 値を最大確率とし，それ以外の Q 値の確率 0 とする．最大の Q 値が 1 つであれば確率 1 とし，複数存在する場合にはすべて同等の確率とする．

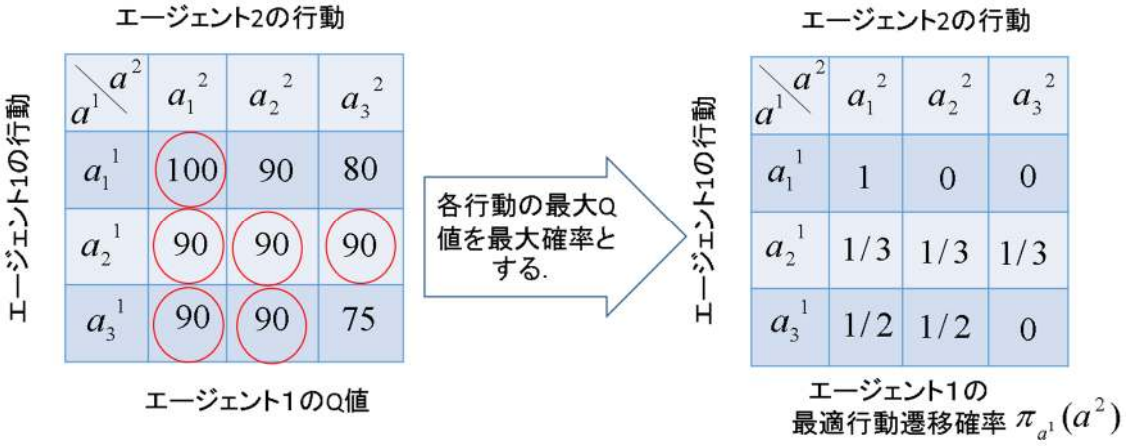


図 12：エージェントの行動遷移確率 $\pi_{a^j}(a^k)$ の計算方法

各エージェント i の各ステップ sp における最適行動 a^i の行動選択方法は Q 学習によって求められた Q 値と Q 値から導かれる行動遷移確率 $\pi_{a^j}(a^k)$ から計算される協調的行動評価値 $r(a^j)$ を元に選択する．協調的行動評価値 $r(a^j)$ の計算方法を図 13 に示す．1 回の出力行動決定の際，始めに各エージェントの行動遷移確率 $\pi_{a^j}(a^k)$ を求める．その後行動選択

の各ステップ sp の際に各エージェントの Q 値と行動遷移確率 $\pi_{a^j}(a^k)$ を元に協調的行動評価値 $r(a^j)$ を求める．協調的行動評価値 $r(a^j)$ は式（1）から求められる． a^j は自エージェントの行動， a^k は他のエージェントの行動， χ は各エージェントの行動番号， AC は各エージェントが所持する行動の数， $Q(S_j, a^j)$ はエージェント j が認識する状態 S_j で他のエージェントが前ステップで行動 a^k を選択した際の行動 a^j の Q 値， $\pi_{a^j}(a^k)$ は自エージェントが行動 a^j を選択していた時に他のエージェントが行動 a^k を選択していた時の行動遷移確率となる．

$$r(a^j) = \sum_{\chi=1}^{AC} Q(S_j, a_x{}^j) \times \pi_{a_\chi{}^j}(a_\delta{}^k) \qquad \cdots (1)$$

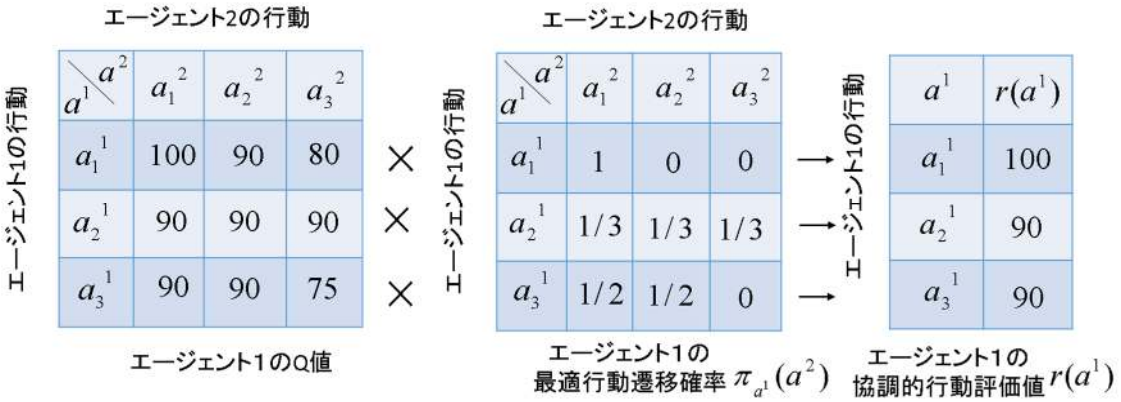


図 13：Q 値と最適行動遷移確率 $\pi_{a^j}(a^k)$ を使用した協調的行動評価値 $r(a^j)$ の計算方法

各エージェントがステップ sp で行動 a^j を選択後，各エージェントが選択した行動 a^j を元に各エージェントが選択した行動 a^j の行動遷移確率 $\pi_{a^j}(a^k)$ を更新する．行動遷移確率 $\pi_{a^j}(a^k)$ の更新方法を図 14 に示す．行動遷移確率 $\pi_{a^j}(a^k)$ の更新方法は式（2），（3）によって行われる．各エージェントが選択した行動 a^j に対して，他のエージェントが選択した行動 a^k に対しては式（2），それ以外の行動に対しては式（3）を適用する． β は定数 $(0 < \beta < 1)$ である．

$$\pi_{a^j}(a^k) \leftarrow \pi_{a^j}(a^k) + \beta\{1 - \pi_{a^j}(a^k)\} \quad \cdots(2)$$

$$\pi_{a^j}(a^k) \leftarrow \pi_{a^j}(a^k) + \beta\{0 - \pi_{a^j}(a^k)\} \quad \cdots(3)$$

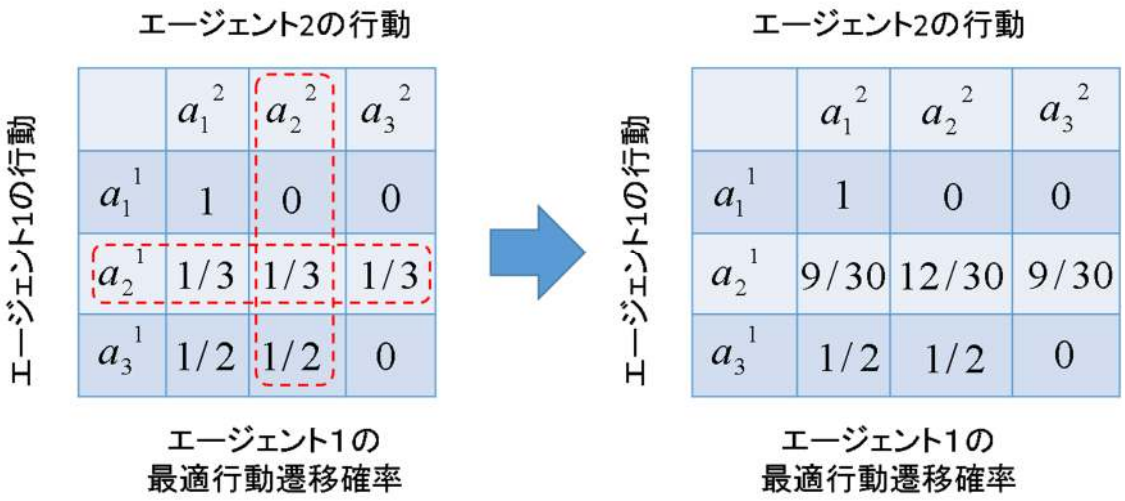


図 14：最適行動遷移確率 $\pi_{a^j}(a^k)$ の更新方法

第 4 章　ロボットアームのリーチング動作による目標物体回収タスク実験

本章では本研究で行ったシミュレーション実験に関する内容について述べる．まず本研究で実験を行う目的を述べる．次に実験で行うタスクとロボット，エージェントの設定について説明をする．そして実験結果を示しこの実験結果から本実験の考察を述べる．

4.1　実験目的

本研究では提案手法の性能を確かめるため，シミュレーションによる従来手法との比較実験を行う．比較内容は提案手法を用いたロボットの行動回収が収束し，従来手法と同等の行動を獲得している点と，本研究の目的であるタスク達成に必要となる行動を学習によって発見し行動回数が収束した後の試行で，最適行動を安定して獲得できるかの 2 点である．本実験での従来手法とは，単体ロボットに対して 1 エージェントによる強化学習を行う手法と，単体ロボットに対して協調動作を学習しないマルチエージェンシステムによる強化学習を行う手法のことである．

4.2　実験概要

4.2.1　タスク設定

本研究ではロボットアームのリーチング動作 [24]による目標物体回収タスクを選択し実験を行った．目標物体回収タスクの概要を図 15 に示す．本実験では実験環境を二次元平面上に表す．ロボットアームには関節を稼働させるサーボモータが搭載されている．またロボットアームの先端には物体を掴むことができるハンドが搭載されている本実験で行うタスクは「ランダムの位置に発生した目標物体をロボットアームのハンドで掴むことで回収する」とする．ロボットアームは各関節を稼働させることでロボットアームの先端を目標物体の位置に合わせる．ロボットアームのハンド部分に目標物体があればハンド部分を稼働させることで目標物体を掴むことができる．

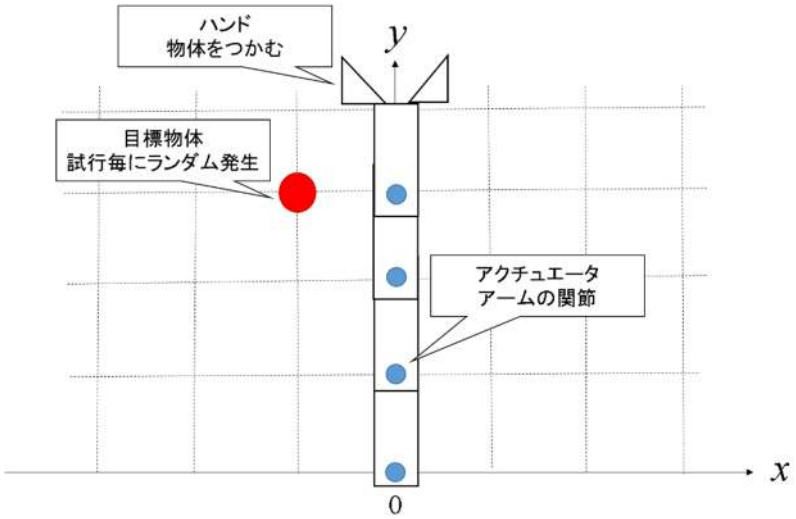


図 15：目標物体回収タスクの概要（アームの関節数が 4 つの場合）

ロボットアームの各関節部分には関節を稼働させるためのサーボモータが搭載されている．ロボットアームの各関節の動作について図 16 に示す．本実験で使用するロボットアームの各関節は 3 方向に稼働する．ロボットアームの各関節は右，正面，左へ稼働の 3 種類である．また各関節の状態も 3 状態となる．なお本実験ではロボットアームの先端のハンド部分の可動はエージェントの行動に含まないことにする．

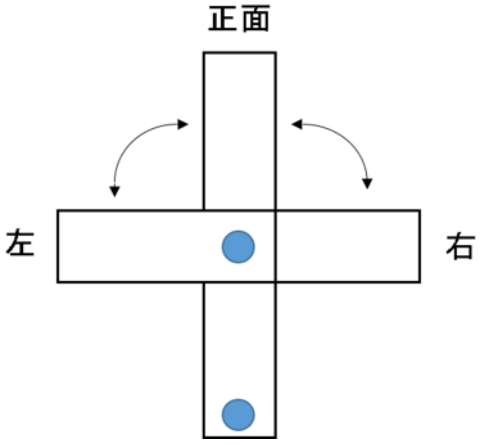


図 16：ロボットアームの各関節の動作

4.2.2　実験設定

本実験では関節数が 2 関節，3 関節，4 関節，5 関節の 4 種類の関節数を所持するロボットアームで実験を行った．ロボットアームの関節数が異なると実験環境の状態数やロボット全体の行動数が異なってくる．そのため状態数や行動数に応じた各手法の結果の違いを比較することができる．

ロボットアームの行動数は各関節の動作が 3 種類となる．そのためロボットアーム全体の行動数は$3 \times$(関節数)となる．ロボットアームが認識する環境状態は目標物体のx座標，y 座標とロボットアームの各関節の状態とする．本実験環境の範囲の設定は表 2 に示す．

表 2：本実験環境の範囲

x 座標の範囲	$\{-1 \times$(関節数 $) \le x \le$(関節数 $)\}$
y 座標の範囲	$\{-1 \times$(関節数 $)+1 \le y \le$(関節数 $)\}$

報酬rは式(4)で決定される．報酬はロボットアームが目標物体を回収したときに$r=100$，それ以外の時は$r=0$となる．

$$r=\begin{cases}100(\text{目標物体を回収したと き}) \\ 0(\text{それ以外})\end{cases} \quad \cdots(4)$$

1 試行はタスクを達成するまで行う．つまりロボットアームが目標物体を回収できるまで行う．1 行動は全アクチュエータの動作とする．また学習を行うタイミングは 1 行動毎に学習を行う．目標物体の発生位置は 1 試行毎にロボットアームの先端が届く位置にランダムに 1 個発生する．ロボットアームの座標上の位置はロボットアームの第一関節部分が二次元座標上の$(0,0)$地点に固定する．アームの初期位置は実験開始時には全関節正面の状態，2 試行目以降では前試行の終了時の状態とする．

本実験での各関節による行動数，目標物体の出現ヵ所の数，環境状態の数を表 3 にまとめる．

表 3：各関節数と行動数，目標物体，環境状態数の対応表

	行動数	目標物体の出現ヵ所の数	環境状態数
2 関節	9	5	45
3 関節	27	10	270
4 関節	81	15	1215
5 関節	243	24	5832

4.2.3 行動学習手法と行動選択手法

本実験では各エージェントの行動学習手法として Q-learning 法 [5]を使用する．Q-learning 法では現時刻tにおいて環境状態S_tで行動a_tを実行した際，報酬r_tを受け取り状態S_{t+1}に遷移したときに式（5）により行動評価値であるQ値を更新する．ただしα $(0<\alpha \le 1)$は学習率，γ $(0 \le \gamma < 1)$は割引値である．

$$Q(S_t, a_t) \leftarrow (1-\alpha)Q(S_t, a_t) + \alpha[r_t \gamma \max_a Q(S_{t+1}, a)] \quad \cdots (5)$$

また本実験では各エージェントの行動選択手法に $\varepsilon - greedy$ 法を使用する．$\varepsilon - greedy$ 法とは定められた確率 ε でランダム行動，$(1-\varepsilon)$ の確率で Q 値の大きい行動を選択する手法である．

4.2.4　シングルエージェントの設定

本実験におけるシングルエージェントとは，ロボットアームの各関節の行動を全て学習し行動選択を行うエージェントのことである．シングルエージェントの場合，行動はロボットアーム全体の行動となる．シングルエージェントの環境状態はロボットアームが認識する目標物体の x 座標，y 座標とロボットアームの各関節の状態となる．

シングルエージェントを適用したロボットアームにおける各関節数に対する 1 エージェントの行動数と状態数の対応について表 4 にまとめる．

表 4：シングルエージェント時の各関節数と 1 エージェントの行動数，状態数の対応表

	1 エージェントの行動数	1 エージェントの環境状態数
2 関節	9	45
3 関節	27	270
4 関節	81	1215
5 関節	243	5832

4.2.5　協調動作を学習しないマルチエージェント

本実験における協調動作を学習しないマルチエージェントとは，ロボットアームの各関節に対してエージェントを設定し，各関節の動作を 1 つのエージェントが学習する手法のことである．本実験では各関節にサーボモータを搭載しているため，関節の数だけエージェントを設定する．各エージェントの行動は各関節の行動 3 種類となる．各エージェントの環境状態はロボットアームが認識する目標物体の x 座標，y 座標とロボットアームの各関節の状態となる．

協調動作を学習しないマルチエージェントを適用したロボットアームにおける各関節数に対するエージェント数と 1 エージェントの行動数と状態数の対応について表 5 にまとめる．

表 5：協調動作を学習しないマルチエージェント時の各関節数とエージェント数と
1 エージェントの行動数，状態数の対応表

	エージェント数	1 エージェントの行動数	1 エージェントの環境状態数
2 関節	2	3	45
3 関節	3	3	270
4 関節	4	3	1215
5 関節	5	3	5832

4.2.6　協調動作を学習するマルチエージェント

協調動作を学習するマルチエージェントとは，本研究での提案手法のことである．協調動作を学習しないマルチエージェントと同様，ロボットアームの各関節にサーボモータを搭載しているため，関節数の数だけエージェントを設定する．各エージェントの行動は各関節の行動 3 種類となる．各エージェントの環境状態ロボットアームが認識する目標物体の x 座標，y 座標とロボットアームの各関節の状態に加え，各エージェントから見た他のエージェントが選択した行動情報が状態となる．

協調動作を学習するマルチエージェントを適用したロボットアームにおける各関節数に対するエージェント数と1エージェントの行動数と状態数の対応について表 6にまとめる．

表 6：協調動作を学習するマルチエージェント時の各関節数とエージェント数と
1 エージェントの行動数，状態数の対応表

	エージェント数	1 エージェントの行動数	1 エージェントの環境状態数
2 関節	2	3	135
3 関節	3	3	2430
4 関節	4	3	32805
5 関節	5	3	472392

4.2.7　実験パラメータ

本研究で行うシミュレーション実験の各種パラメータを表 7 に示す．ロボットアームの行動は 1 行動で全ての関節を稼働し，目標物体を回収するまでを 1 試行とする．目標物体を回収したとき，次の試行に移る．

表 7：実験パラメータ

試行回数	200000 （回）
タスク達成報酬	$r=100$
ε	0.05
N	100
ステップサイズ・パラメータα	0.1
割引値γ	0.9
β	0.01

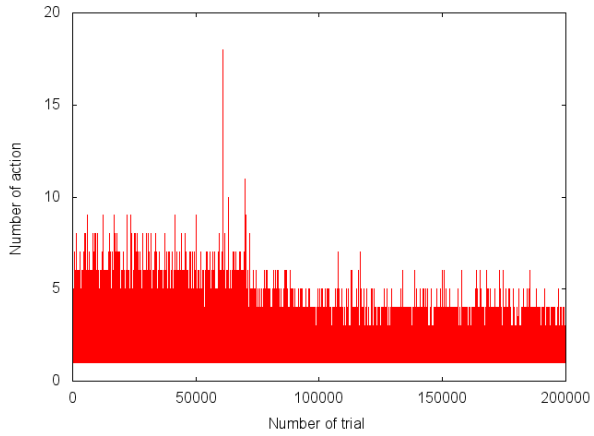
4.3 実験結果

本節では前節の設定で行ったシミュレーション実験の結果を示し，その結果について説明する．本実験では関節数が 2 関節，3 関節，4 関節，5 関節のロボットアームに対して，シングルエージェント，協調動作を学習しないマルチエージェント，提案手法を適用して実験を行った．提示する実験結果は，各試行での行動数の推移，3 試行間毎の行動数の平均値の推移，各試行時点での累計行動数の推移，各試行時点での経験済みの状態行動対の数と割合の 4 種類である．それぞれの実験結果はシングルエージェントの場合，協調動作を学習しないマルチエージェントの場合，提案手法の場合で結果を示し，その結果について手法ごとに比較する．

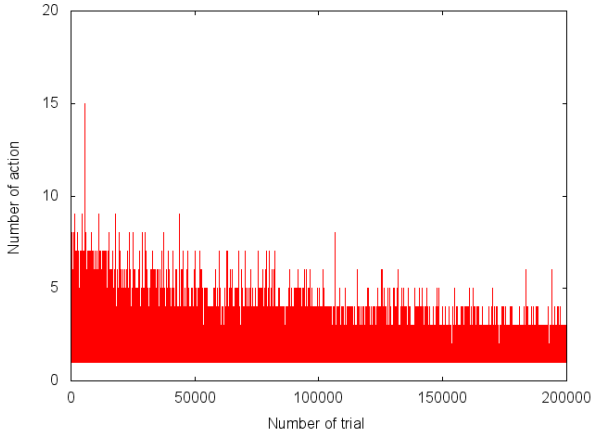
4.3.1 2 関節ロボットアームの場合

本節では 2 関節のロボットアームによるリーチング動作による目標物体回収タスクの実験結果を示す。

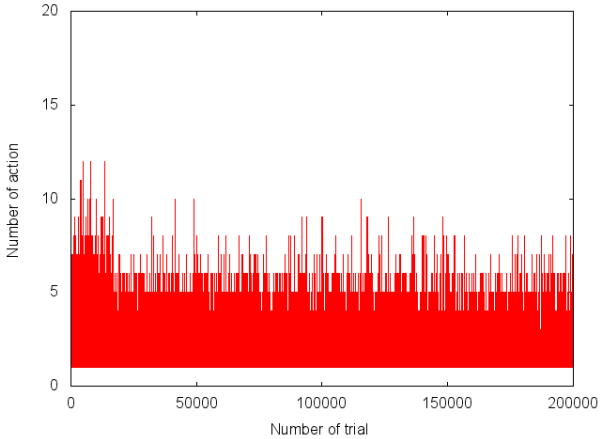
始めに 3 手法それぞれの各試行での行動数の推移を図 17，図 18，図 19 に示す．横軸は試行数，縦軸は 1 試行の行動数である．今回の実験設定では 1 回の行動で目標物体を回収することが可能な設定にしている．そのため 1 試行の行動数が 1 に収束していれば行動数が収束し，学習が完了しているといえる．図 19 から 3 手法共に行動回数が 1 に収束していることが分かる．したがって 3 手法共に学習によってタスク達成に必要となる行動を獲得していることが分かる．この結果から提案手法は試行を繰り返す中で，学習を行いシングルエージェントや協調動作を学習しないマルチエージェントと同等の行動を獲得することが示された．しかし一方で試行間の行動回数の変移を見ると，提案手法は従来手法 2 種よりも行動回数が多くなる試行が多くなっていることが読み取れる．これは協調動作を行わなくても最適行動を獲得することが可能であることが示される．ただし 2 関節のロボットアームでは行動数や目標物体の数も少ないので協調動作を学習しないマルチエージェントでも十分にタスク達成に必要な行動を選択可能であるとも考えられる．



(a)：シングルエージェント

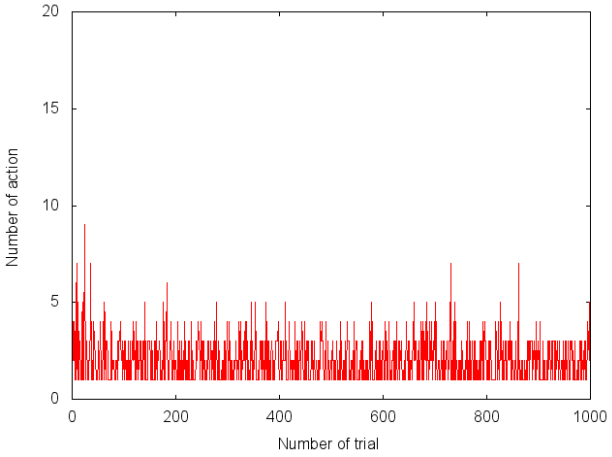


(b)：協調動作を学習しないマルチエージェント

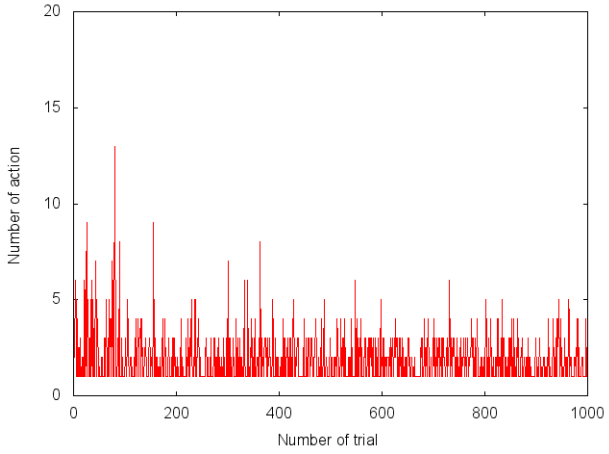


(c)：提案手法

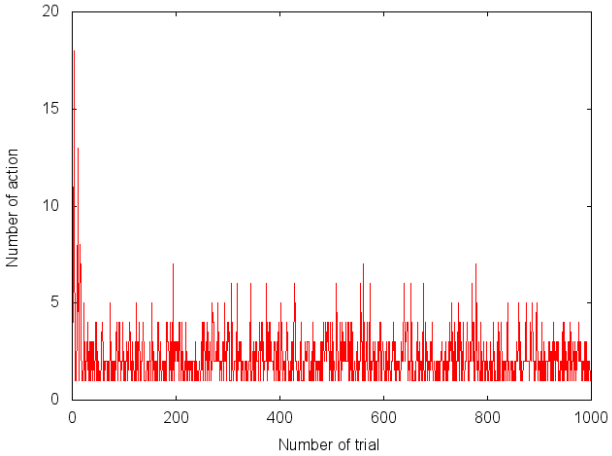
図 17：2 関節時の各手法の各試行での行動数の推移



(a)：シングルエージェント

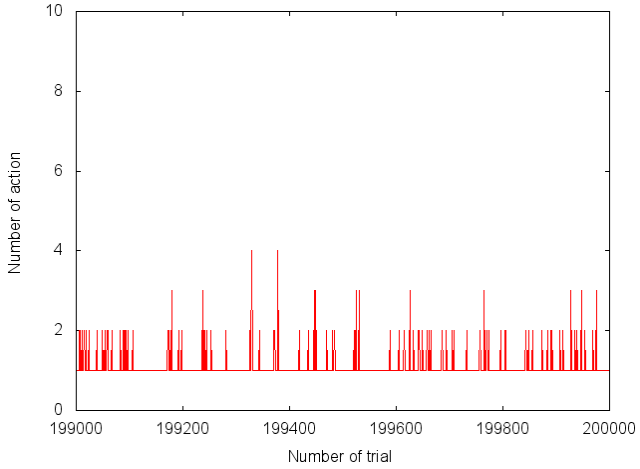


(b)：協調動作を学習しないマルチエージェント

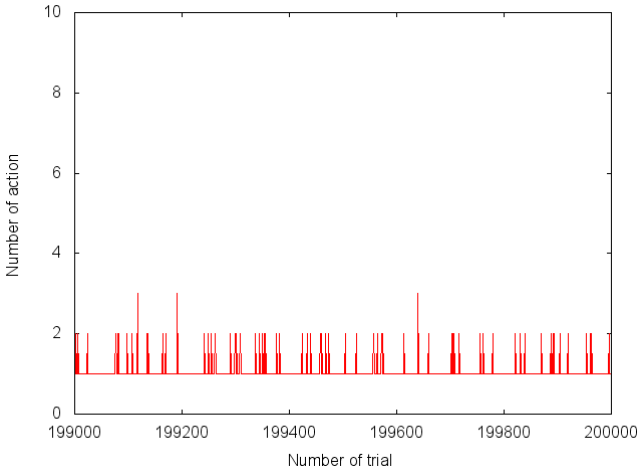


(c)：提案手法

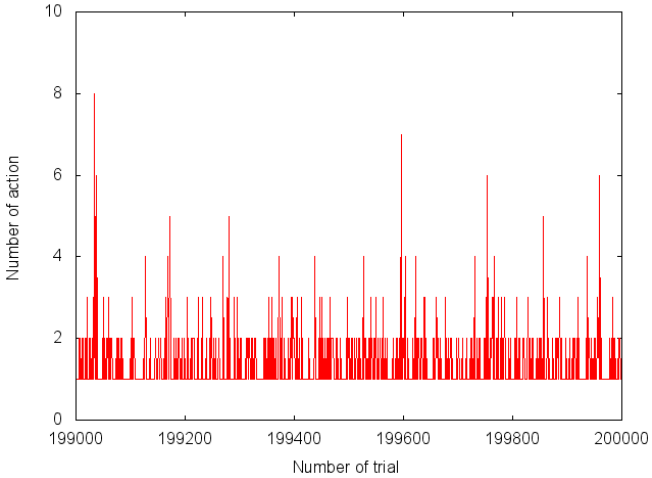
図 18：2 関節時の各手法の各試行での行動数の推移（1 試行から 1000 試行）



(a)：シングルエージェント



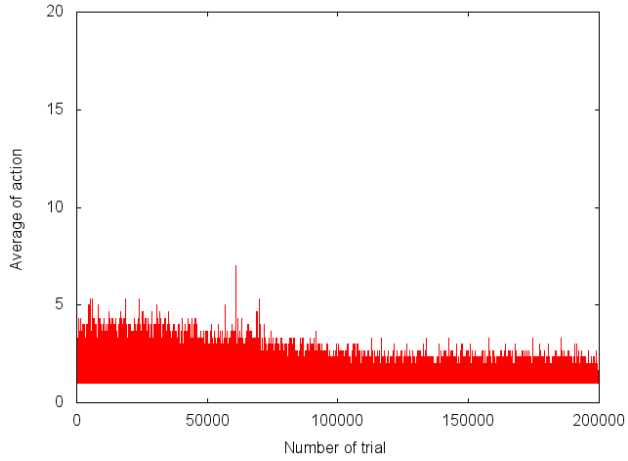
(b)：協調動作を学習しないマルチエージェント



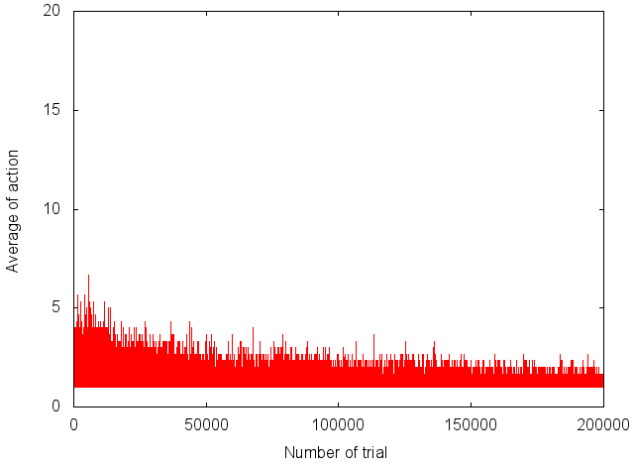
(c)：提案手法

図 19：2 関節時の各手法の各試行での行動数の推移（199000 試行から 200000 試行）

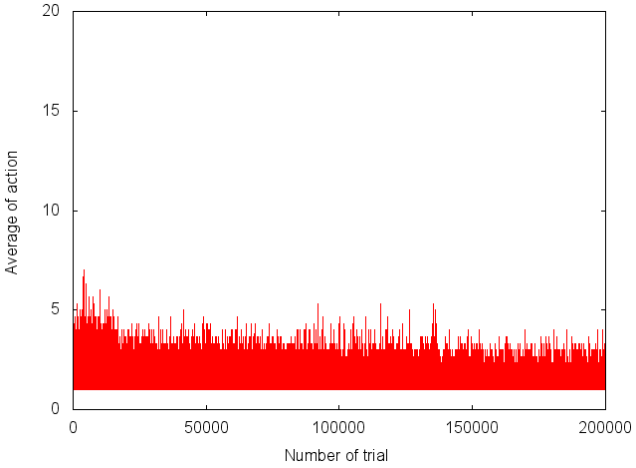
次に各手法の 3 試行の行動数の平均値の推移を図 20，図 21，図 22 に示す．横軸は試行数，縦軸は 3 試行の行動数の平均値となる．今回の実験では 1 回の行動で目標物体を回収することが可能な設定にしている．そのため 3 試行の行動数の平均値が 1 に収束していれば行動数が収束し学習が完了しているといえる．図 22 を見ると，提案手法の行動が収束すると行動回数が 1 となっていることが分かる．このことから提案手法を用いたロボットは学習によってタスク達成に必要となる行動を獲得していることが分かる．この結果から提案手法は試行を繰り返す中で学習を行い，シングルエージェントや協調動作を学習しないマルチエージェントと同等の行動を獲得することが示された，しかし一方で試行間の行動数の変異を見ると提案手法は従来手法 2 種よりも行動回数が多くなる試行が多くなっていることが読み取れる．これは協調動作を行わなくても最適行動を獲得することが可能であることが示される．ただし 2 関節のロボットアームの場合では行動数や目標物体の数も少ないので協調動作を学習しないマルチエージェントでも十分にタスク達成に必要な行動を選択可能であるとも考えられる．



(a)：シングルエージェント

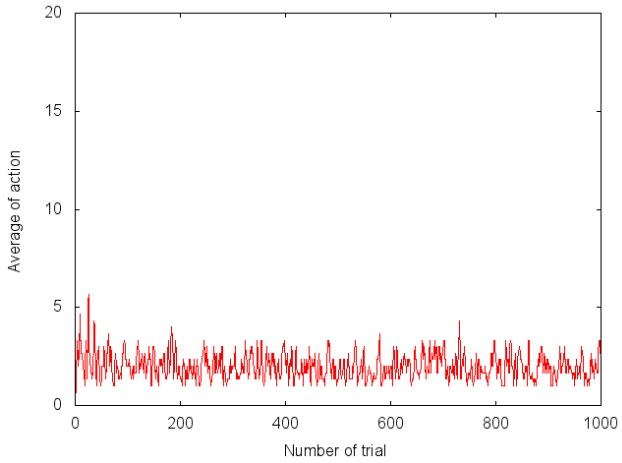


(b)：協調動作を学習しないマルチエージェント

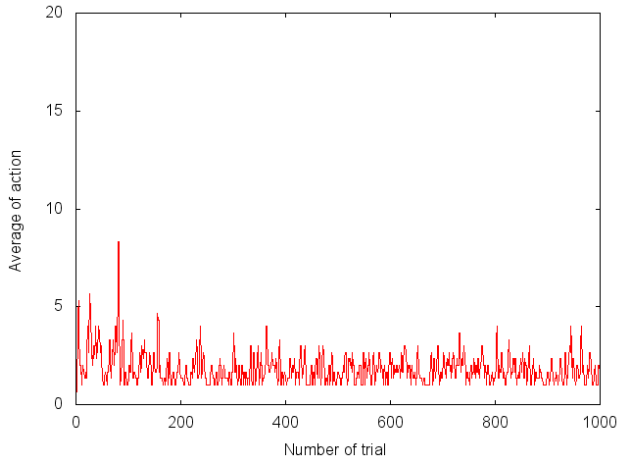


(c)：提案手法

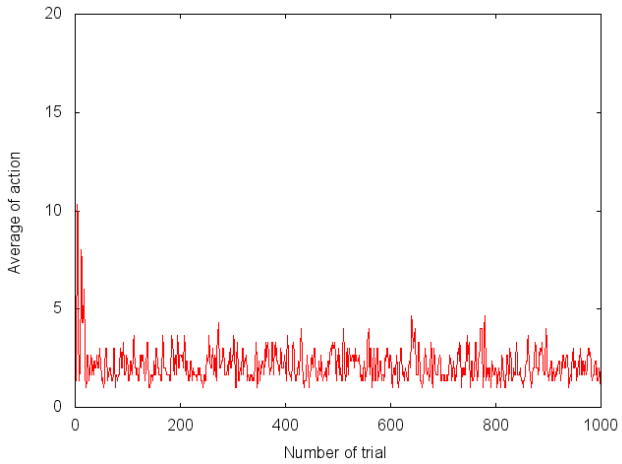
図 20：2 関節時の各手法の 3 試行での行動数の平均値の推移



(a)：シングルエージェント

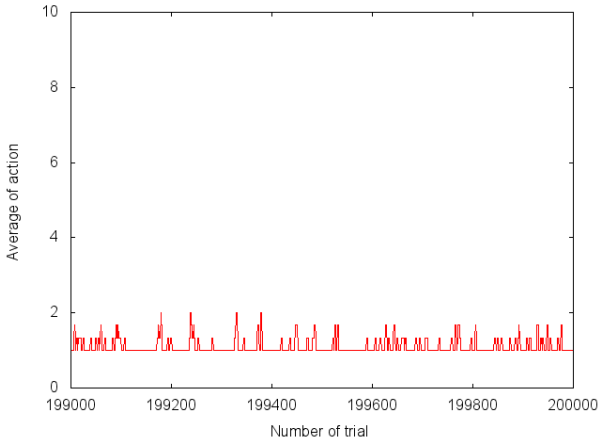


(b)：協調動作を学習しないマルチエージェント

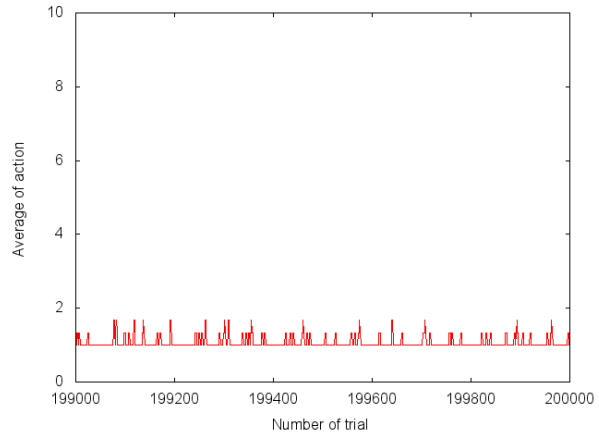


(c)：提案手法

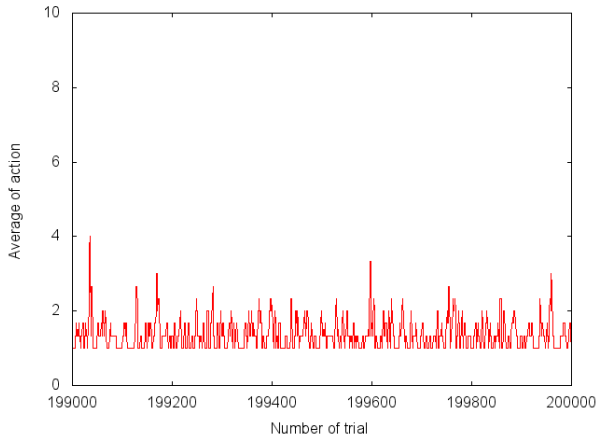
図 21：2 関節時の各手法の 3 試行での行動数の平均値の推移（1 試行から 1000 試行）



(a)：シングルエージェント



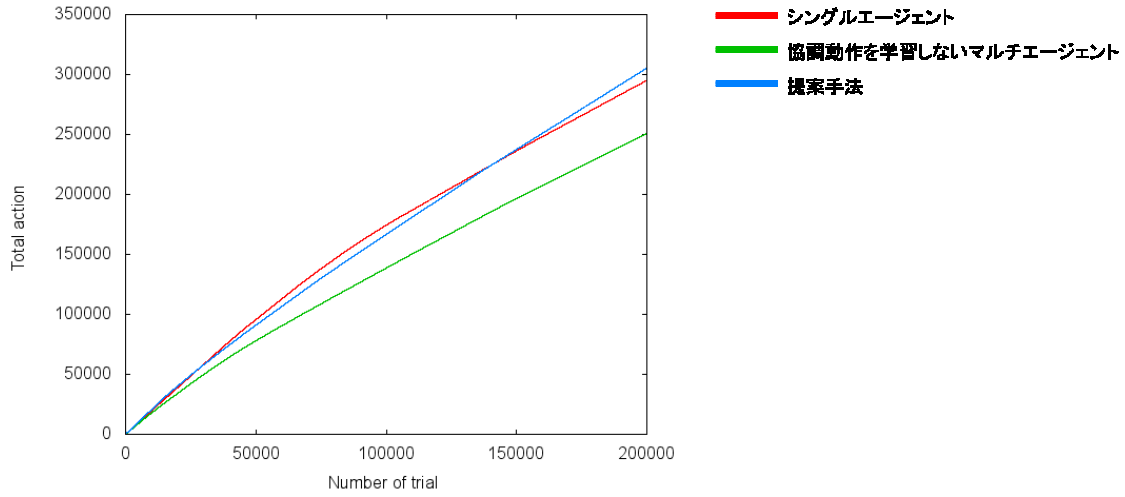
(b)：協調動作を学習しないマルチエージェント



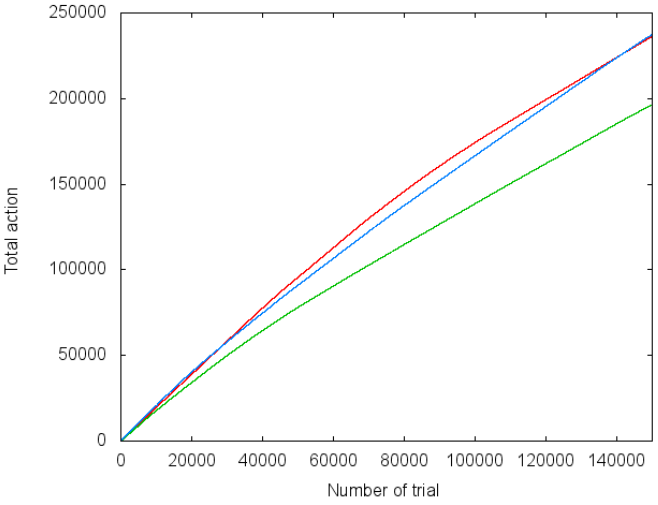
(c)：提案手法

図 22：2 関節時の各手法の 3 試行での行動数の平均値の推移（199000 試行から 200000 試行）

次に各手法の各試行時点での 1 試行目からの累計行動数の推移を図 23 に示す．横軸は試行回数，縦軸は各試行時点での累計行動数である．図 23 から提案手法を用いたエージェントは最終的に一番行動数が多くなっていることが分かる．その理由として 2 関節時には状態数，行動数が少ない．そのため エージェント間の協調行動が必要となる場面が少ない．したがって協調動作を学習しないマルチエージェントでもタスク達成に必要となる行動を選択することが可能であるため提案手法の有効性が働いてこないと考えられる．提案手法は協調動作を学習しないマルチエージェントよりも 1 エージェントの状態数が多いため学習が収束するまでの試行数が多くなる．さらに提案手法は各エージェントが探査的行動を行うか判定している．そのためロボット全体の行動の中にランダムによって選択された行動が含まれる確率はεの値より高くなる．そのためシングルエージェントよりもランダム行動を選択する確率が高くなる．これらの理由から提案手法の累計行動数は従来手法 2 種よりも多くなると考えられる．



(a)：1 試行から 200000 試行



(b)：1 試行から 150000 試行

図 23：2 関節時の各手法の各試行時点での累計行動数の推移

次に各手法の各試行時点での経験済みの状態行動対の数と割合を図 24，図 25 に示す．横軸は試行数，縦軸は図 24 では各試行時点での経験済みの状態行動対の数，図 25 では各試行時点での経験済みの状態行動対の割合を示す．図 25 を見ると 2 手法共に 8 割以上の状態行動対を経験していることが分かる．この結果から 3 手法共にタスク達成に対して十分な状態行動対を経験している状態にあるといえる．

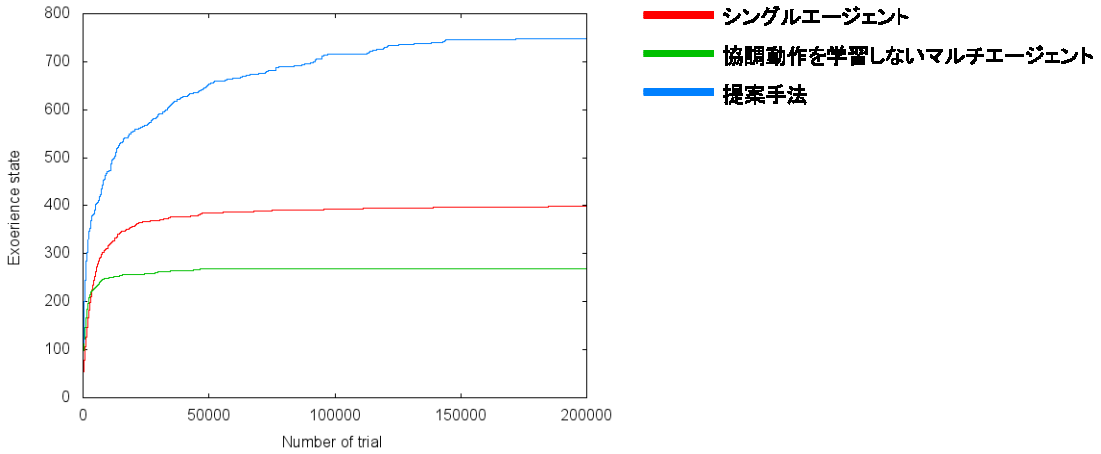


図 24：2 関節時の各手法の各試行時点での経験済み状態行動対の数

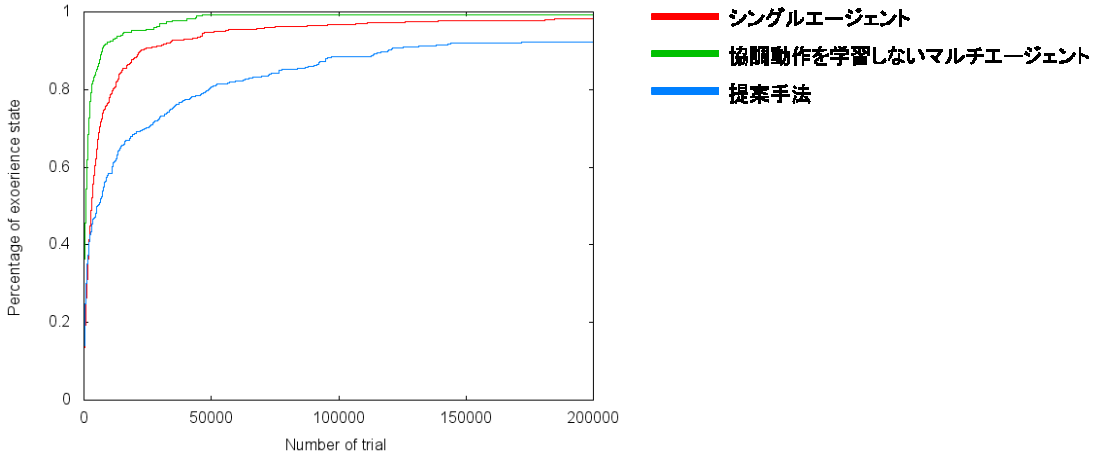
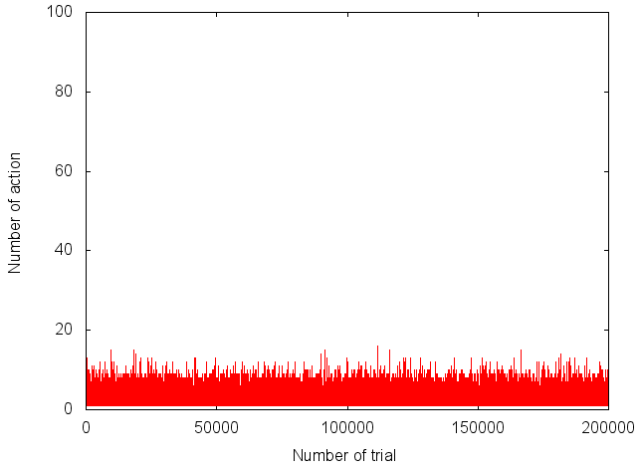


図 25：2 関節時の各手法の各試行時点での経験済み状態行動対の割合

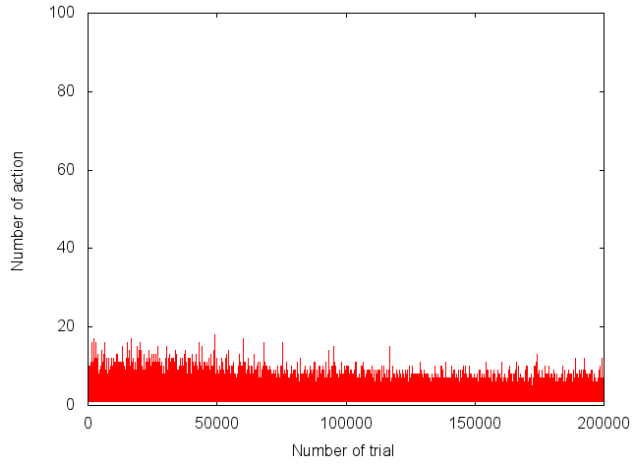
4.3.2　3 関節ロボットアームの場合

本節では 3 関節のロボットアームによるリーチング動作による目標物体回収タスクの実験結果を示す.

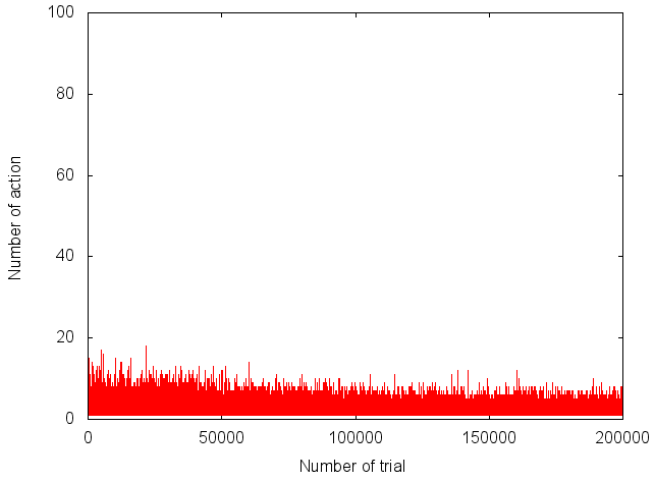
始めに 3 手法それぞれの各試行での行動数の推移を図 26, 図 27, 図 28 に示す. 横軸は試行数, 縦軸は 1 試行の行動数である. 今回の実験設定では 1 回の行動で目標物体を回収することが可能な設定にしている. そのため 1 試行の行動数が 1 に収束していれば行動数が収束し学習が完了しているといえる. 図 28 から 3 手法共に行動回数が 1 に収束していることが分かる. したがって 3 手法共に学習によってタスク達成に必要となる行動を獲得していることが分かる. この結果から提案手法は試行を繰り返す中で学習を行い, シングルエージェントや協調動作を行わないマルチエージェントと同等の行動を獲得していることが示された. また行動数の変移を見ると提案手法は従来手法 2 種よりも少ない行動数であることが分かる. これはエージェントの一部がランダムに行動を選択しても残りのエージェントがランダム行動を選択したエージェントに合わせた行動選択を行っているため未経験の状態行動対に遷移しにくいためと考えられる. このことから提案手法はエージェント間の協調行動を学習してタスク達成するために必要となる行動を選択できていることが示された.



(a)：シングルエージェント

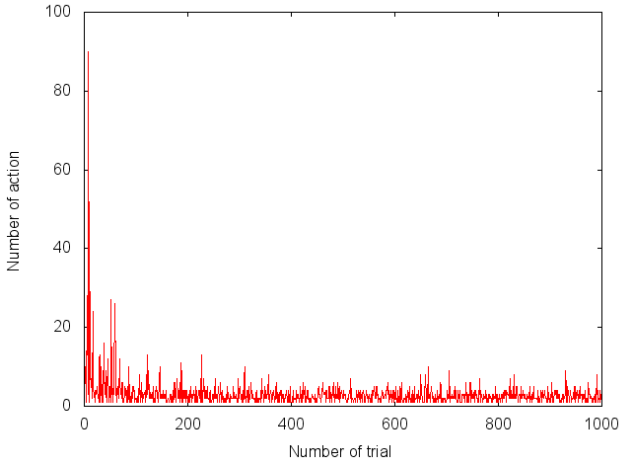


(b)：協調動作を学習しないマルチエージェント

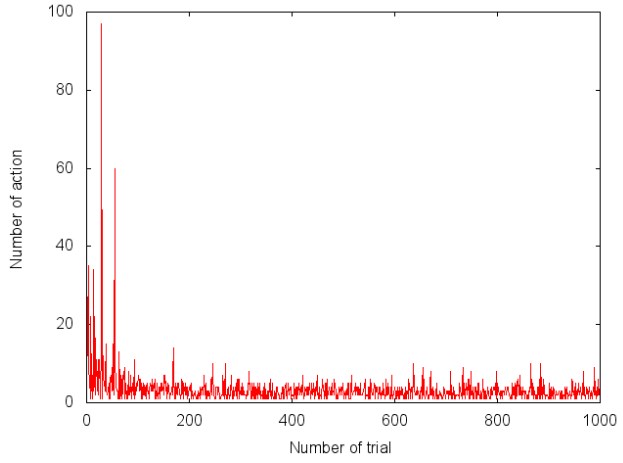


(c)：提案手法

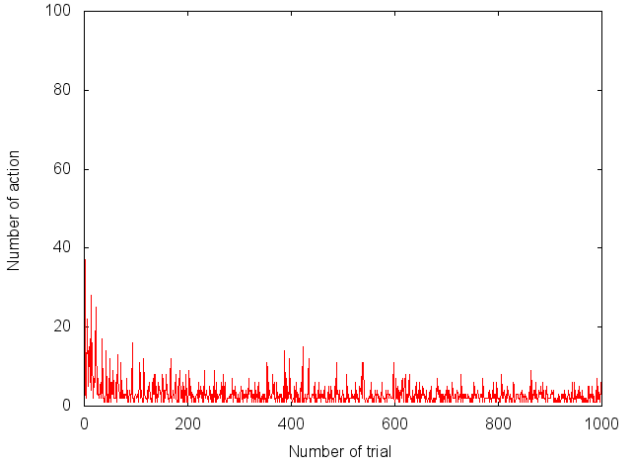
図 26：3 関節時の各手法の各試行での行動数の推移



(a)：シングルエージェント

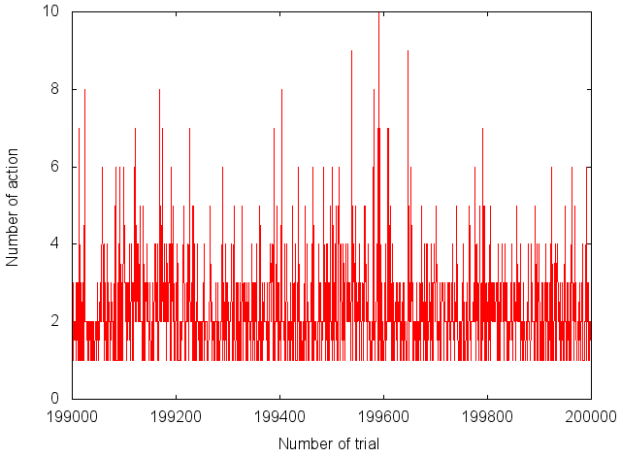


(b)：協調動作を学習しないマルチエージェント

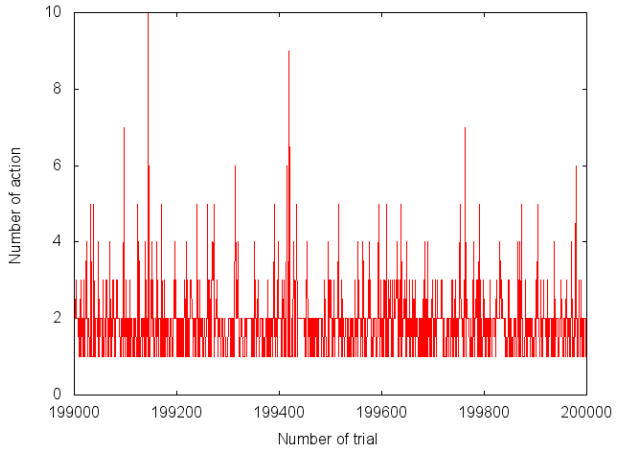


(c)：提案手法

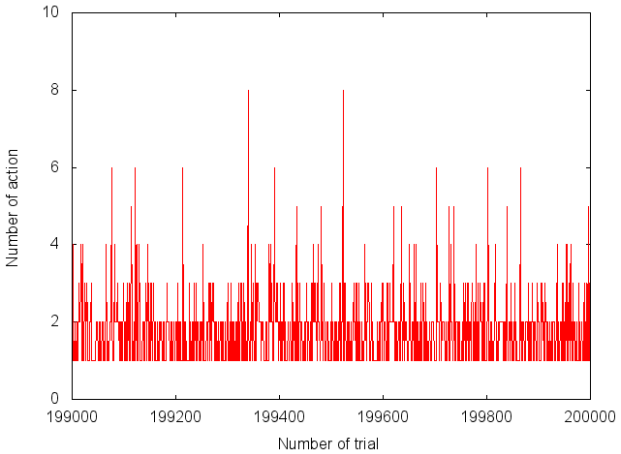
図 27：3 関節時の各手法の各試行での行動数の推移（1 試行から 1000 試行）



(a)：シングルエージェント



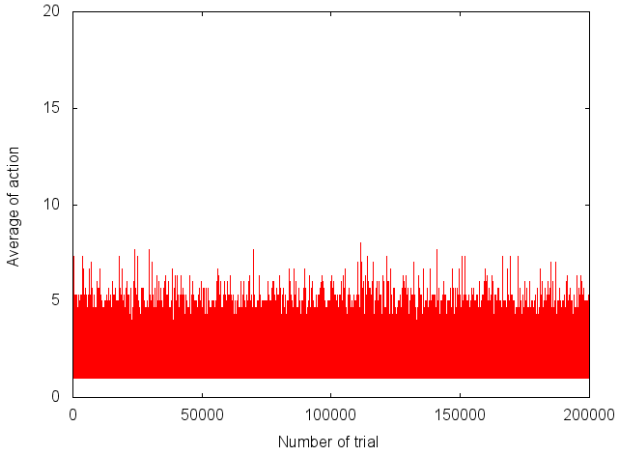
(b)：協調動作を学習しないマルチエージェント



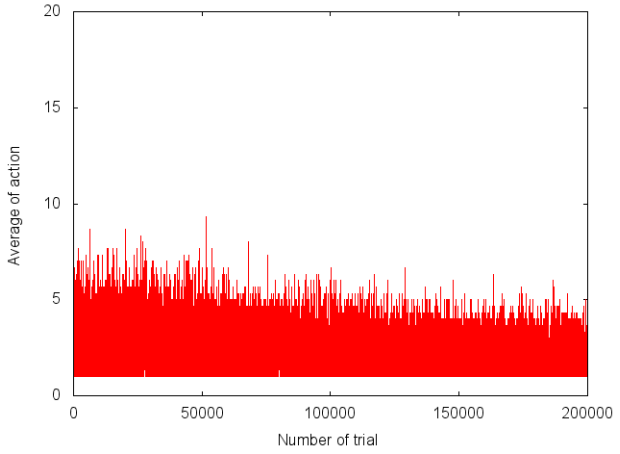
(c)：提案手法

図 28：3 関節時の各手法の各試行での行動数の推移（199000 試行から 200000 試行）

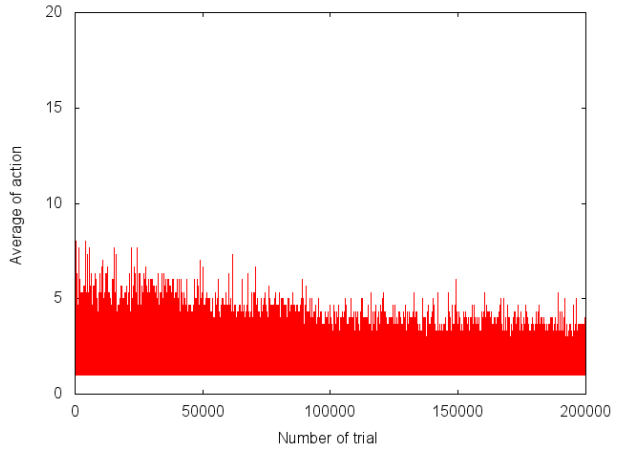
次に各手法の 3 試行の行動数の平均値の推移を図 29，図 30，図 31 に示す．横軸は試行数，縦軸は 3 試行の行動数の平均値となる．今回の実験では 1 回の行動で目標物体を回収することが可能な設定にしている．そのため 3 試行の行動数の平均値が 1 に収束していれば行動数が収束し学習が完了しているといえる. 図 31 を見ると提案手法の行動数が収束すると行動数が 1 となっていることが分かる．このことから提案手法を用いたロボットは学習によってタスク達成に必要となる行動を獲得していることが分かる．この結果から提案手法は試行を繰り返す中で学習を行い，シングルエージェントや協調動作を学習しないマルチエージェントと同等の行動を獲得することが示された．また行動数の平均値の変移を見ると提案手法は従来手法 2 種よりも少ない平均値であることが分かる．これはエージェントの一部がランダムに行動を選択しても残りのエージェントがランダム行動を選択したエージェントに合わせた行動選択を行っているため未経験の状態行動対に遷移しにくいためと考えられる．このことから提案手法はエージェント間の協調行動を学習してタスク達成するために必要となる行動を選択できていることが示された.



(a)：シングルエージェント

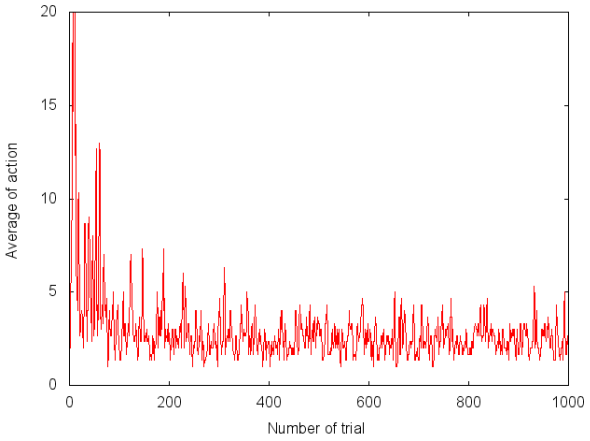


(b)：協調動作を学習しないマルチエージェント

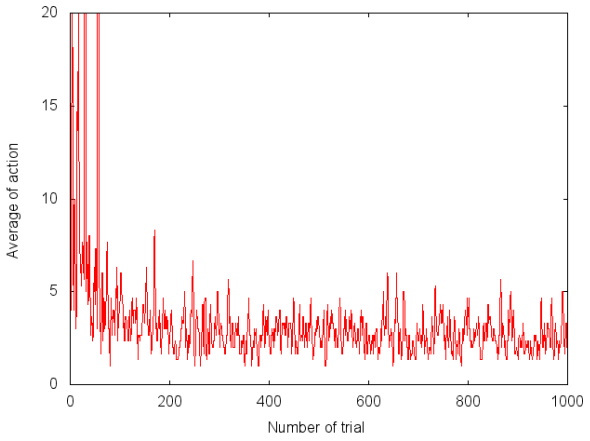


(c)：提案手法

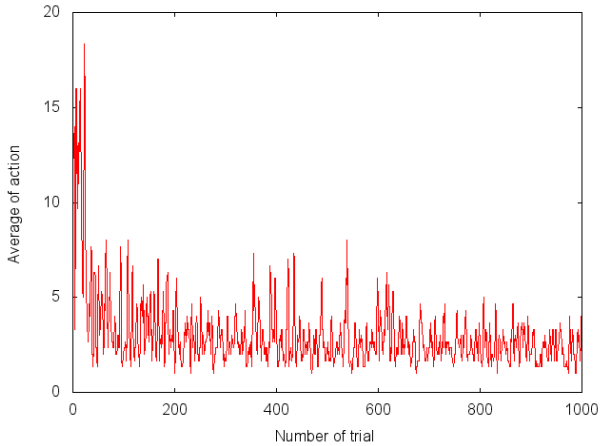
図 29：3 関節時の各手法の 3 試行での行動数の平均値の推移



(a)：シングルエージェント

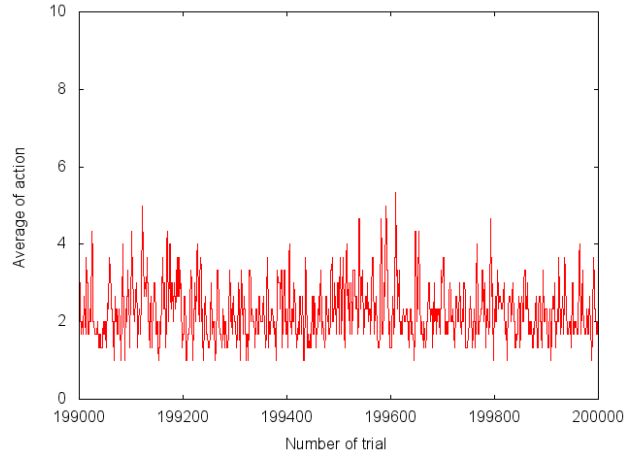


(b)：協調動作を学習しないマルチエージェント

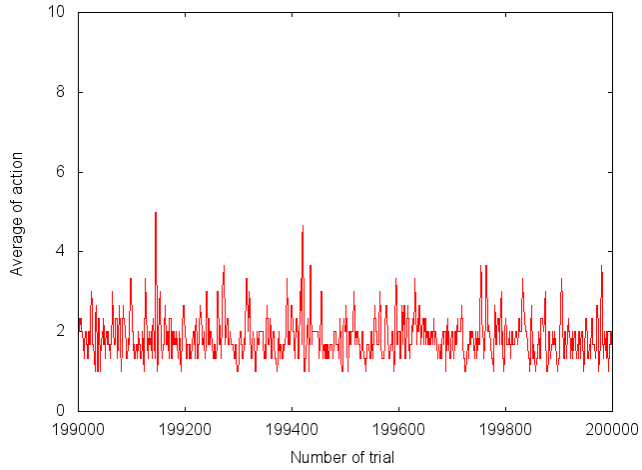


(c)：提案手法

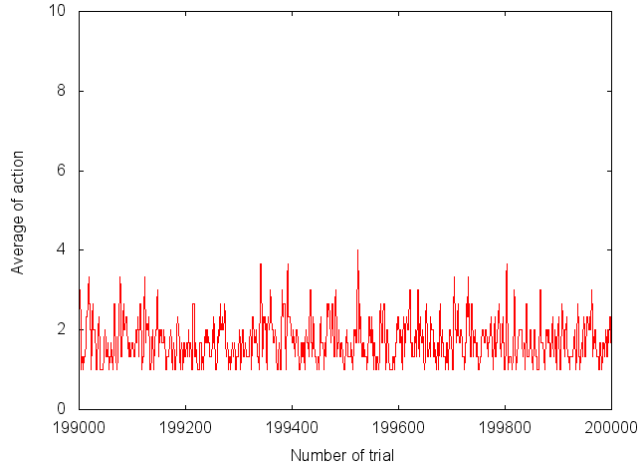
図 30：3 関節時の各手法の 3 試行での行動数の平均値の推移（1 試行から 1000 試行）



(a)：シングルエージェント



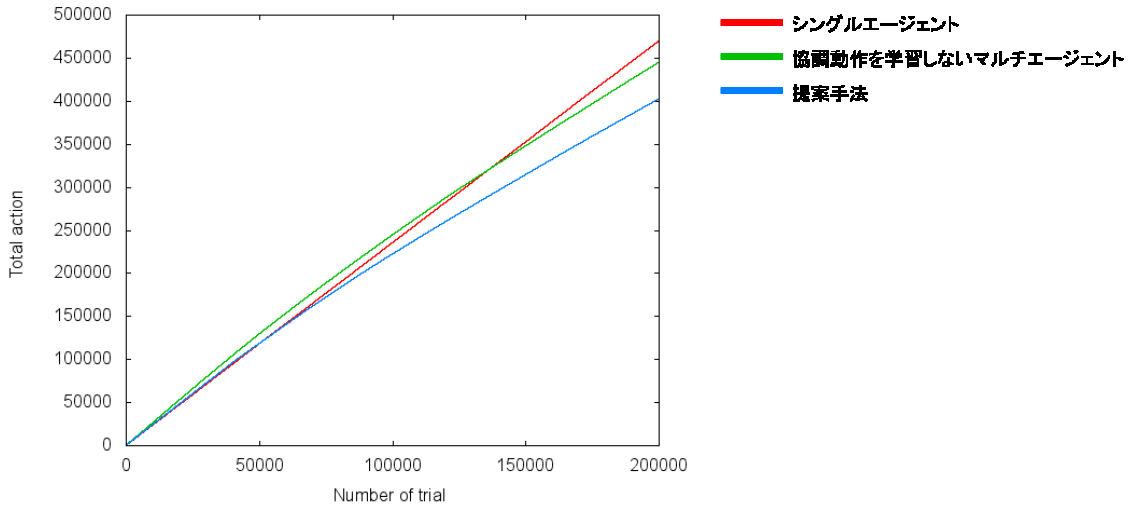
(b)：協調動作を学習しないマルチエージェント



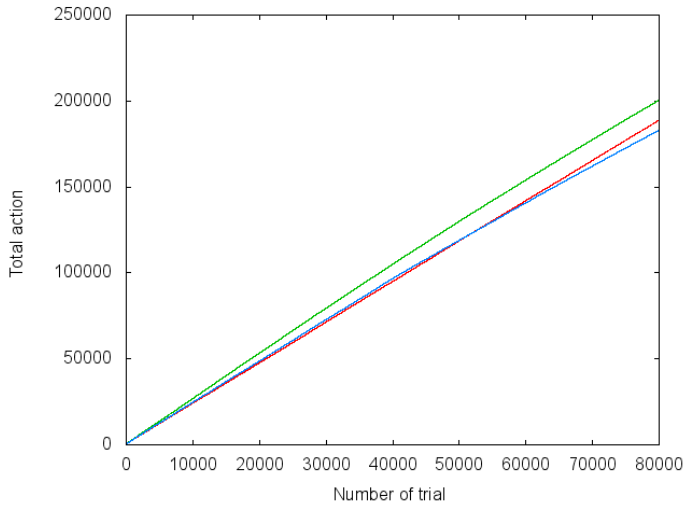
(c)：提案手法

図 31：3 関節時の各手法の 3 試行での行動数の平均値の推移（199000 試行から 200000 試行）

次に各手法の各試行時点での 1 試行目からの累計行動数を図 32 に示す．横軸は試行回数，縦軸は各試行時点での累計行動数である．図 32 から提案手法の行動数は従来手法 2 種より下回っていることが分かる．この理由として提案手法はエージェント間の協調行動を学習しているため 1 つの目標物体に対して複数の最適行動がある時に，複数の行動から 1 つの行動を選択することができる。一方で協調動作を学習しないマルチエージェントの場合では複数の最適行動の内，1 つの行動を選択することができない．そのため提案手法は確実に最適行動を選択できるため累計行動数が少なくなると考えられる．シングルエージェントの場合では 1 つの状態に対する行動が多くなる．そのため 1 つの状態に対する最適行動を発見するまでの試行数が増加する．そのためマルチエージェント手法よりも学習が遅れてしまうことになると考えられる．この結果から提案手法はエージェント間の協調動作を学習しタスク達成に必要となる行動を獲得していることが示された．



（a）：1 試行から 200000 試行



（b）：1 試行から 80000 試行

図 32：3 関節時の各手法の各試行時点での累計行動数の推移

次に各手法の各試行時点での経験済みの状態行動対の数と割合を図 33，図 34 に示す．横軸は試行数，縦軸は図 33 では各試行時点での経験済みの状態行動対の数，図 34 では各試行時点での経験済みの状態行動対の割合を示す．図 34 を見ると協調動作を学習しないマルチエージェントはほぼすべての状態行動対を経験している．一方でシングルエージェントと提案手法は約 5 割の状態行動対を経験していることが分かる．この結果からシングルエージェントと提案手法に関しては 200000 試行以上行うことでロボットの行動がよくなる可能性があるが，協調動作を学習しないマルチエージェントに関しては 200000 試行以上行ったとしてもこれ以上ロボットの行動が良くなることはないと考えられる．

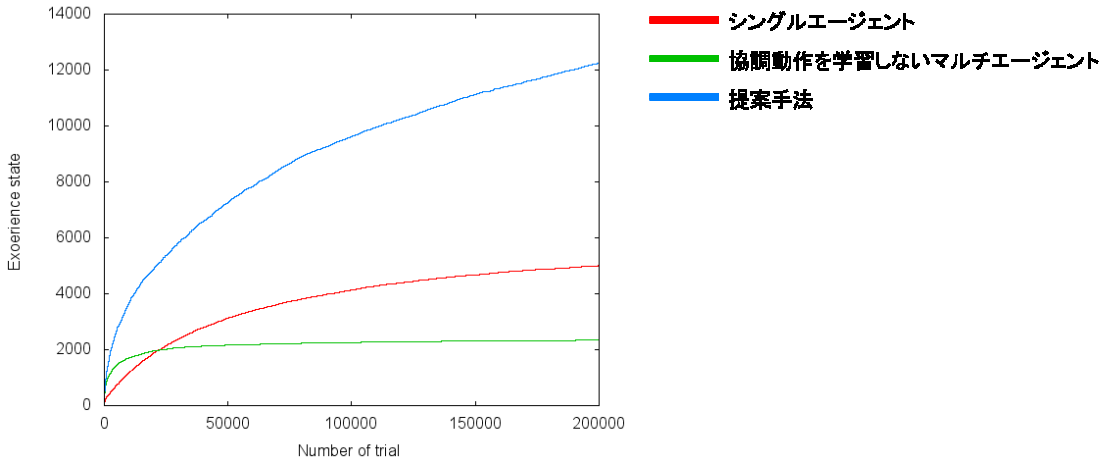


図 33：3 関節時の各手法の各試行時点での経験済み状態行動対の数

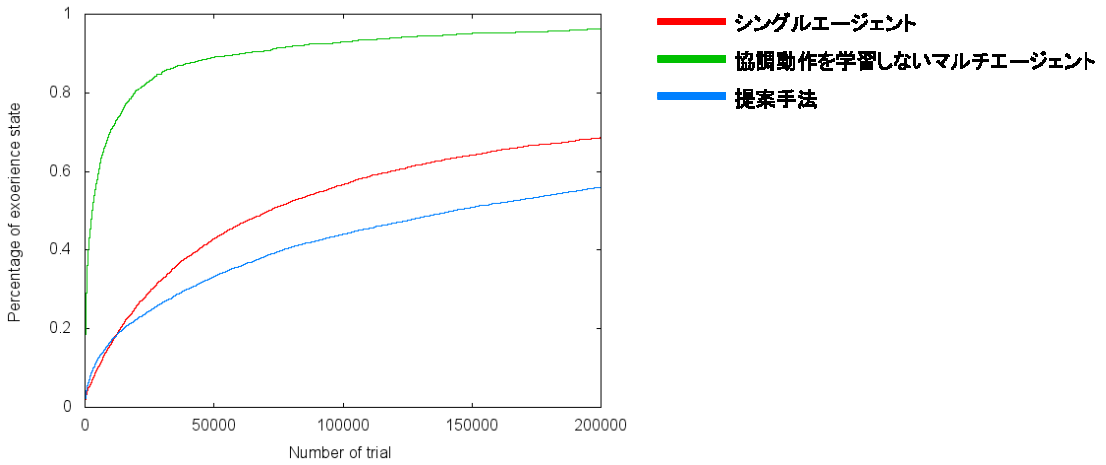
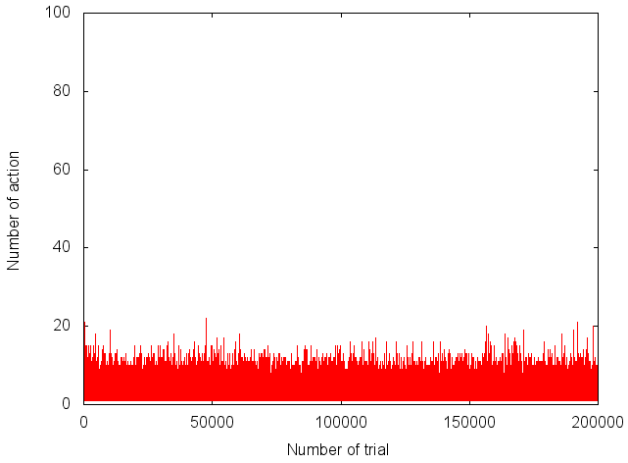


図 34：3 関節時の各手法の各試行時点での経験済み状態行動対の割合

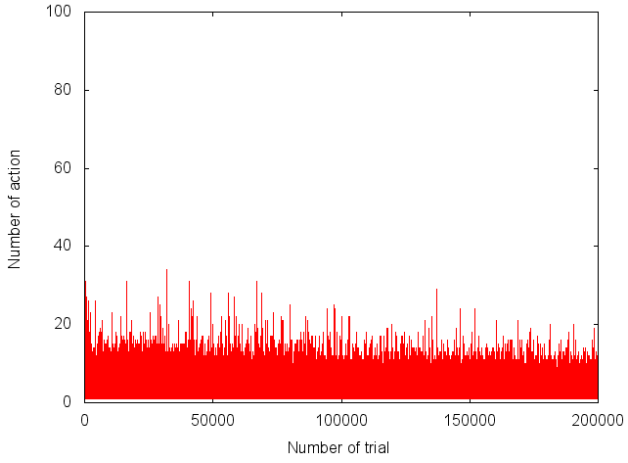
4.3.3　4 関節ロボットアームの場合

本節では 4 関節のロボットアームによるリーチング動作による目標物体回収タスクの実験結果を示す.

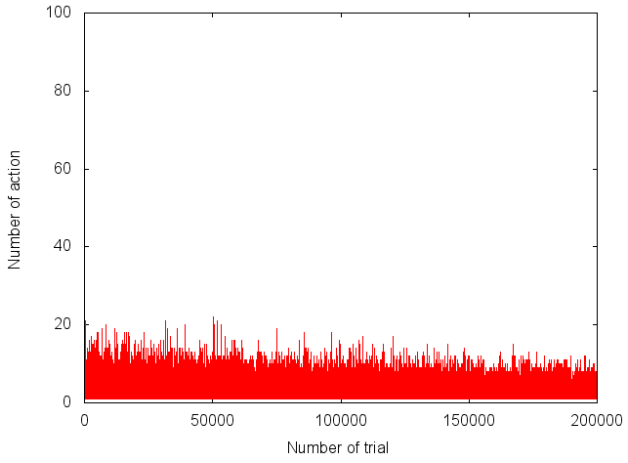
始めに 3 手法それぞれの各試行での行動数の推移を図 35, 図 36, 図 37 に示す. 横軸は試行数, 縦軸は 1 試行の行動数である. 今回の実験設定では 1 回の行動で目標物体を回収することが可能な設定にしている. そのため 1 試行の行動数が 1 に収束していれば行動数が収束し学習が完了しているといえる. 図 37 から 3 手法共に行動数が 1 付近で収束していることが分かる. したがって提案手法は従来手法 2 種と同等の行動を獲得していることが分かる. この結果から提案手法は試行を繰り返す中で学習を行い, シングルエージェントや協調動作を学習しないマルチエージェントと同等の行動を獲得することが示された.



(a)：シングルエージェント

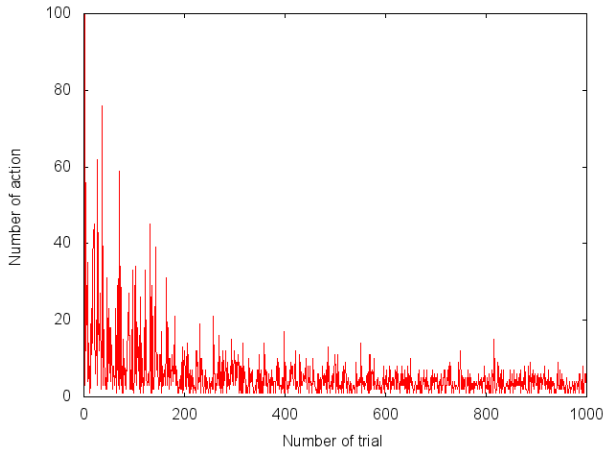


(b)：協調動作を学習しないマルチエージェント

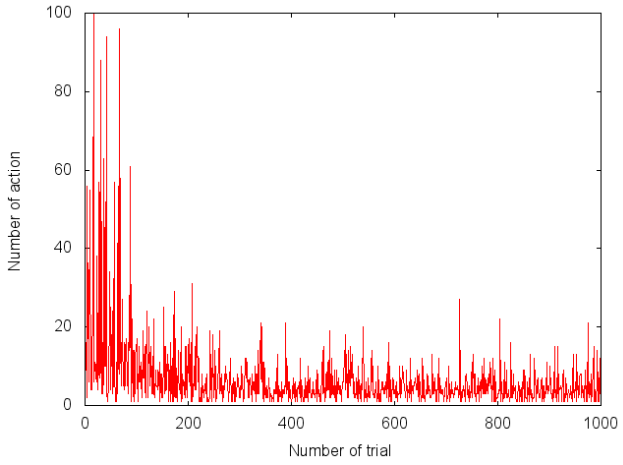


(c)：提案手法

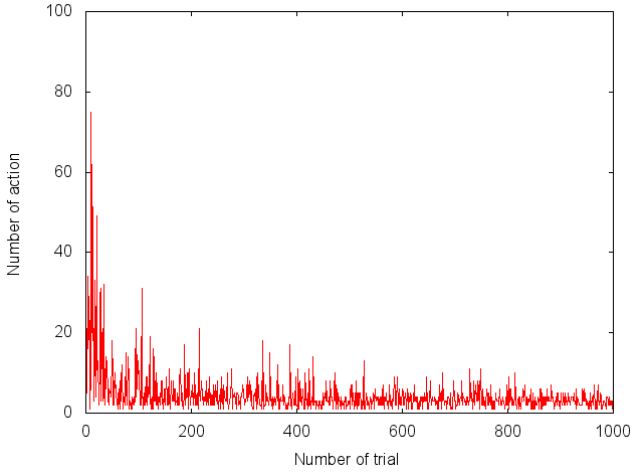
図 35：4 関節時の各手法の各試行での行動数の推移



(a)：シングルエージェント

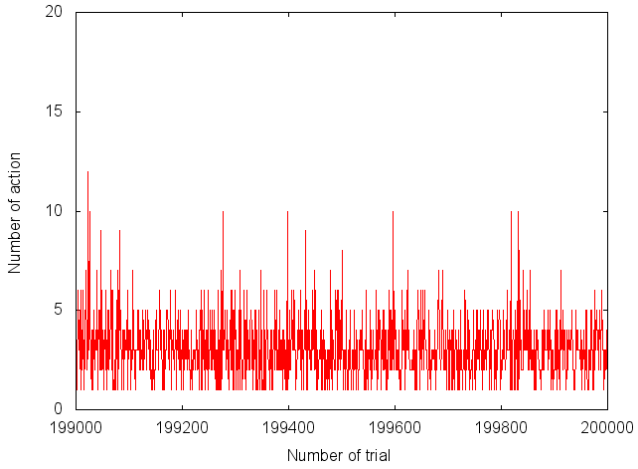


(b)：協調動作を学習しないマルチエージェント

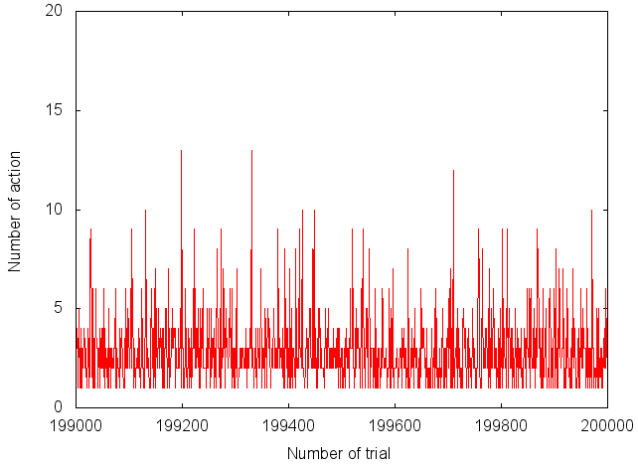


(c)：提案手法

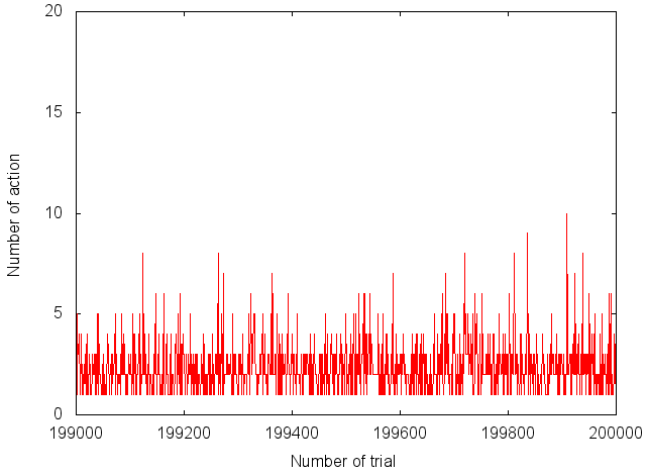
図 36：4 関節時の各手法の各試行での行動数の推移（1 試行から 1000 試行）



(a)：シングルエージェント



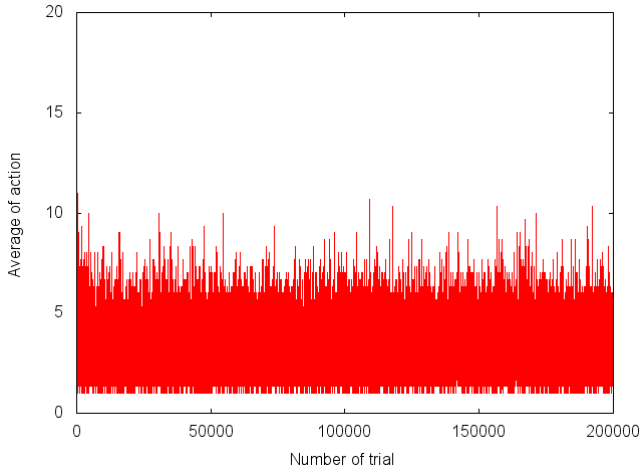
(b)：協調動作を学習しないマルチエージェント



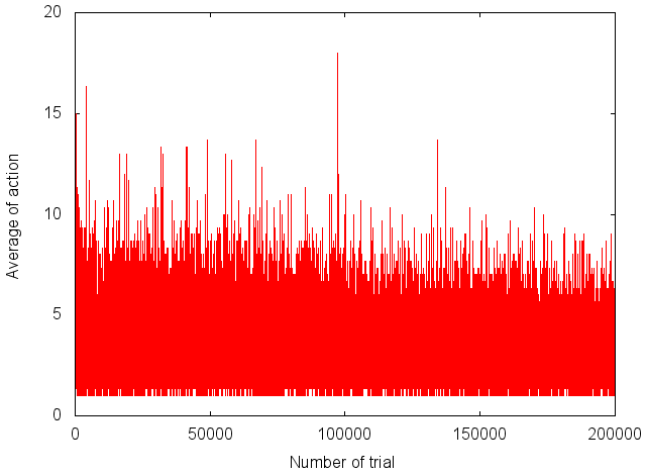
(c)：提案手法

図 37：4 関節時の各手法の各試行での行動数の推移（199000 試行から 200000 試行）

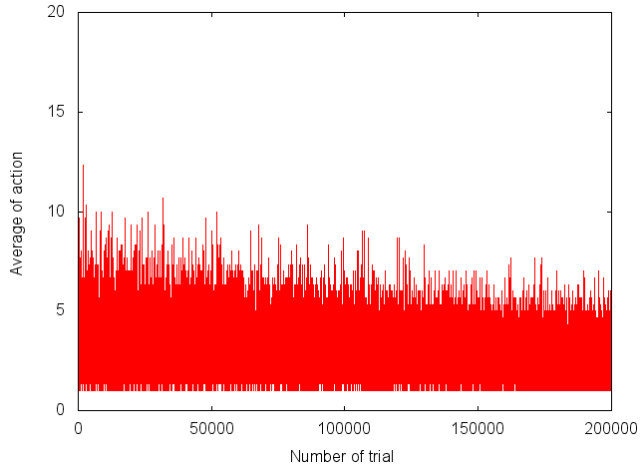
次に各手法の 3 試行の行動数の平均値の推移を図 38，図 39，図 40 に示す．横軸は試行数，縦軸は 3 試行の行動数の平均値となる．今回の実験では 1 回の行動で目標物体を回収することが可能な設定にしている．そのため 3 試行の行動数の平均値が 1 に収束していれば行動数が収束し学習が完了しているといえる．図 40 から 3 試行共に行動数の平均値が 1 に収束していることが分かる．したがって提案手法は従来手法 2 種と同等の行動を獲得していることが分かる．この結果から提案手法は試行を繰り返す中で学習を行い，シングルエージェントや協調動作を学習しないマルチエージェントと同等の行動を獲得することが示された．



(a)：シングルエージェント

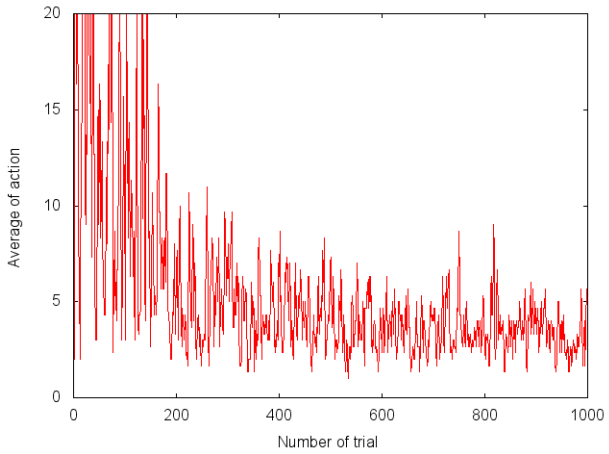


(b)：協調動作を学習しないマルチエージェント

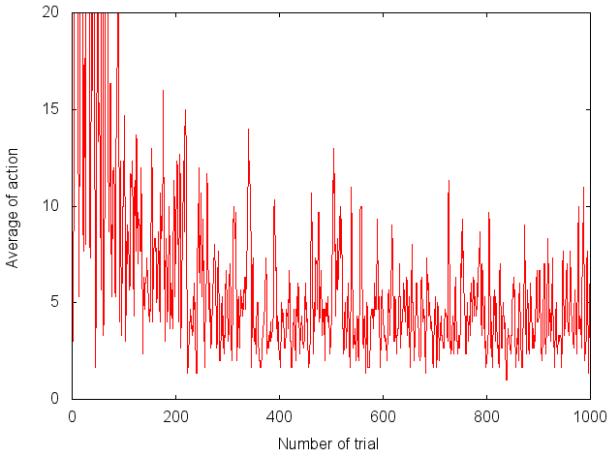


(c)：提案手法

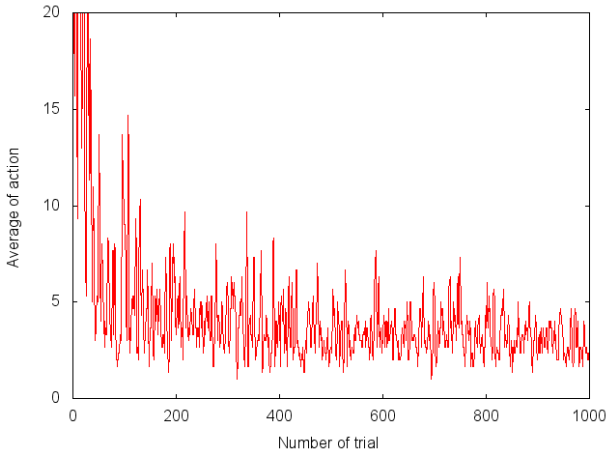
図 38：4 関節時の各手法の 3 試行での行動数の平均値の推移



(a)：シングルエージェント

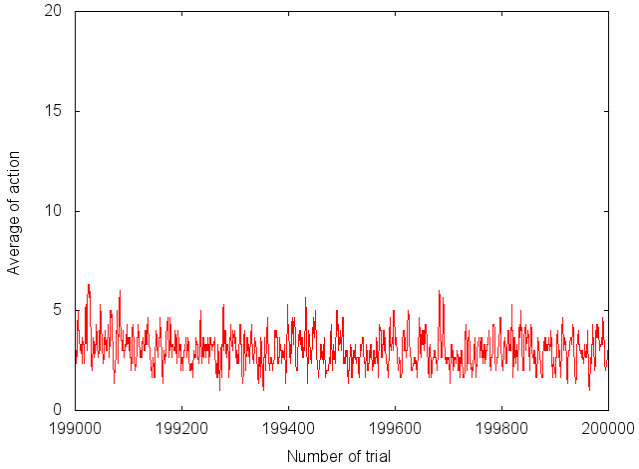


(b)：協調動作を学習しないマルチエージェント

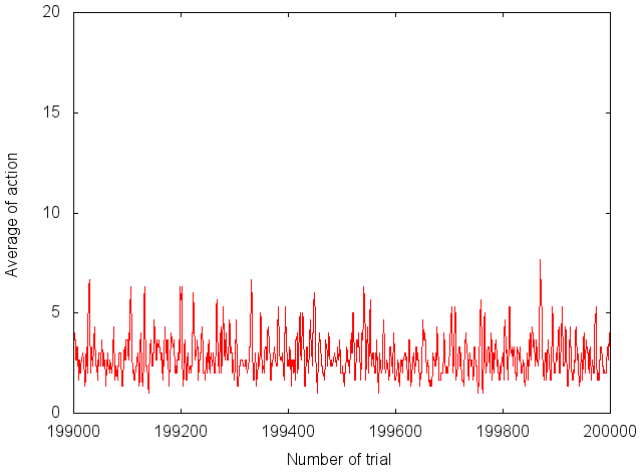


(c)：提案手法

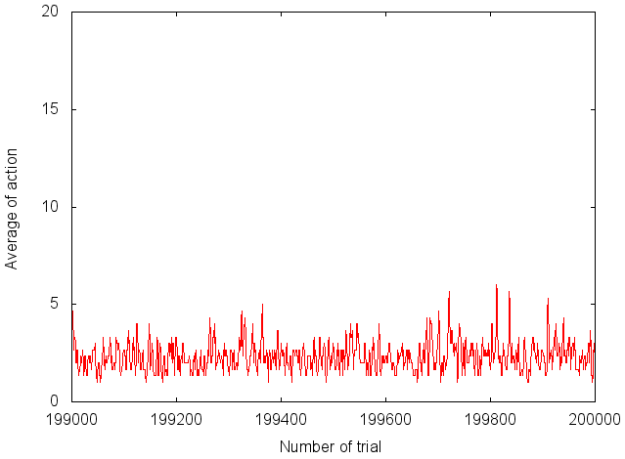
図 39：4 関節時の各手法の 3 試行での行動数の平均値の推移（1 試行から 1000 試行）



(a)：シングルエージェント



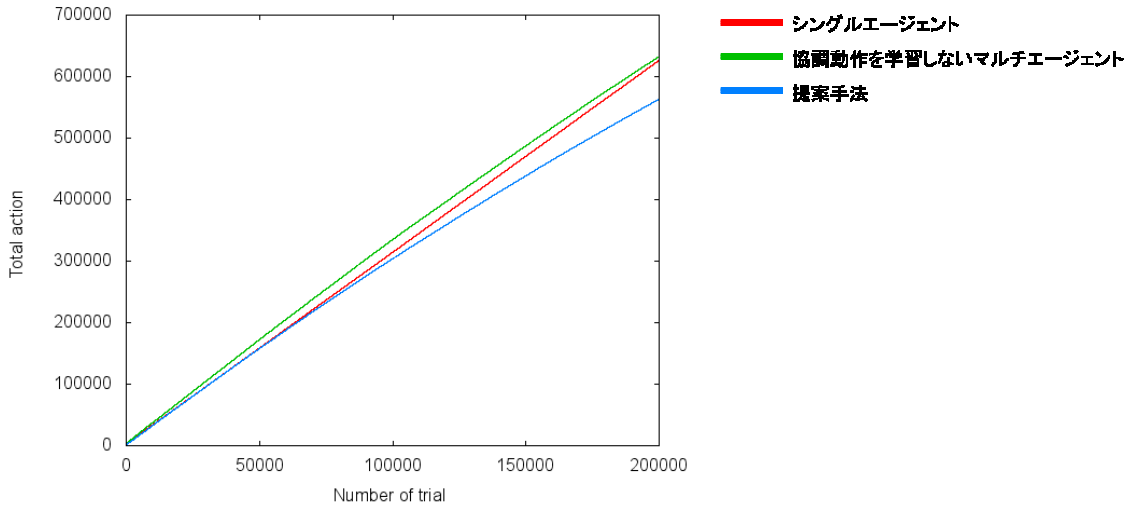
(b)：協調動作を学習しないマルチエージェント



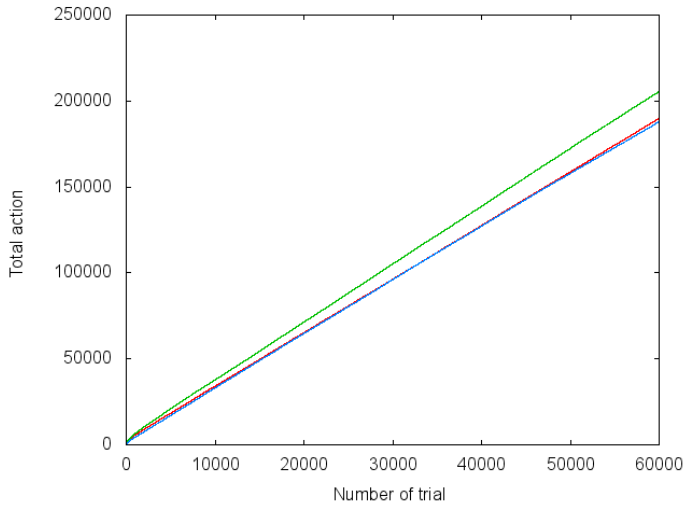
(c)：提案手法

図 40：4 関節時の各手法の 3 試行での行動数の平均値の推移（199000 試行から 200000 試行）

次に各手法の各試行時点での 1 試行目からの累計行動数を図 41 に示す．横軸は試行数，縦軸は各試行時点での累計行動数である．図 41 から提案手法の行動数の累計は従来手法 2 種と比較して少なくなっていることが示された．この理由としてロボットアームの関節数が 4 つに増加したことによってロボットの状態行動対が増加し行動回数が収束するまでの試行数が増加していると考えられる．その結果 3 手法共に行動数が完全に 1 に収束していない状態となっていると考えられる．また関節数が増加したことで最適行動が複数存在する目標物体の位置の数も増加している．そのため協調動作を学習しないマルチエージェントを用いたロボットアームは各エージェントの行動学習だけではロボットの最適行動を選択することが困難となっている．その結果協調動作を学習しないマルチエージェントはタスクを達成するまでの行動数が増加していると考えられる．一方で提案手法はエージェント間の協調動作を獲得しているため最適行動が複数存在する場合でもロボットの最適行動を選択できる．そのため提案手法の行動数が協調なしマルチエージェントの行動数を下回っていると考えられる．



（a）：1 試行から 200000 試行



（b）：1 試行から 60000 試行

図 41：4 関節時の各手法の各試行時点での累計行動数の推移

次に各手法の各試行時点での経験済みの状態行動対の数と割合を図 42，図 43 に示す．横軸は試行数，縦軸は図 42 では各試行時点での経験済みの状態行動対の数，図 43 では各試行時点での経験済みの状態行動対の割合を示す．図 43 を見ると協調動作を学習しないマルチエージェントは 8 割以上の状態行動対を経験している．一方でシングルエージェントと提案手法は約 2 割の状態行動対しか経験していないことが分かる．この結果から協調動作を学習しないマルチエージェントは 200000 試行行うことで環境に対して十分な状態行動対を経験していることが分かる．一方でシングルエージェントと提案手法に関しては 200000 試行では十分な状態行動対を経験していないことが分かる．しかし 200000 試行時点のペースから予測すると，十分な状態行動対を経験するためにはかなりの試行数が必要になると考えられる．

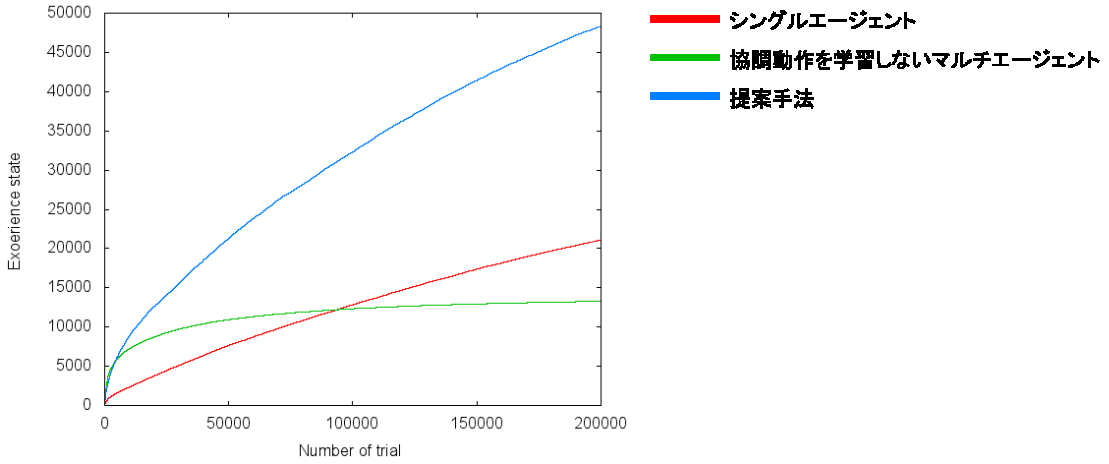


図 42：4 関節時の各手法の各試行時点での経験済み状態行動対の数

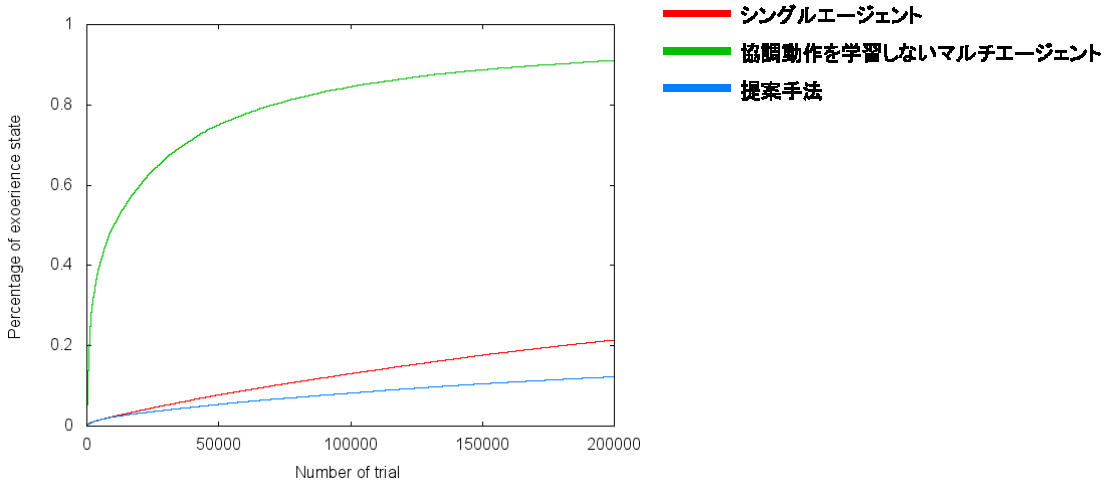
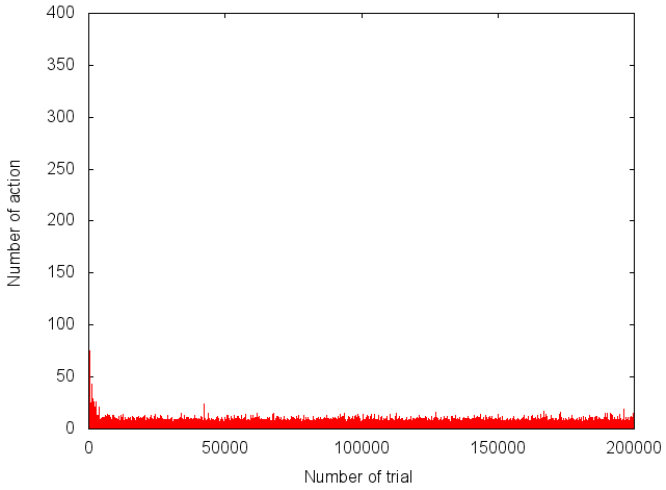


図 43：4 関節時の各手法の各試行時点での経験済み状態行動対の割合

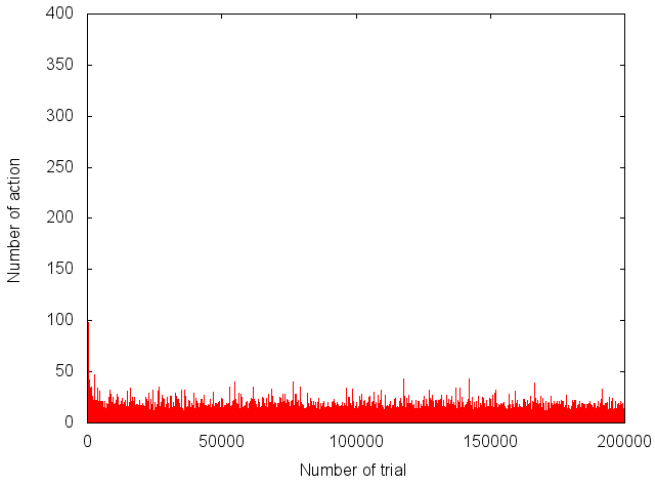
4.3.4　5 関節ロボットアームの場合

本節では 5 関節のロボットアームのリーチング動作による目標物体回収タスクの実験結果を示す.

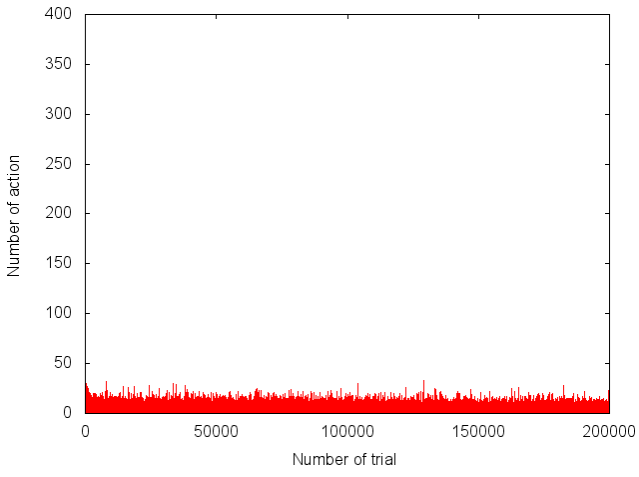
始めに 3 手法それぞれの各試行時点での行動数の推移を図 44，図 45，図 46 に示す. 横軸は試行数，縦軸は 1 試行の行動数である．今回の実験設定では 1 回の行動で目標物体を回収することが可能な設定にしている．そのため 1 試行の行動数が 1 に収束していれば行動数が収束し学習が完了しているといえる．図 46 から 3 手法共に行動数が 1 付近で収束していることが分かる．この結果から提案手法は試行を繰り返す中で学習を行い，シングルエージェントや協調動作を学習しないマルチエージェントと同等の行動を獲得することが示された.



(a)：シングルエージェント

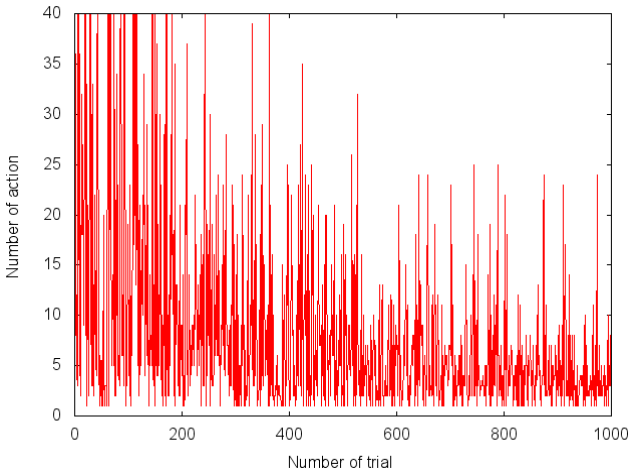


(b)：協調動作を学習しないマルチエージェント

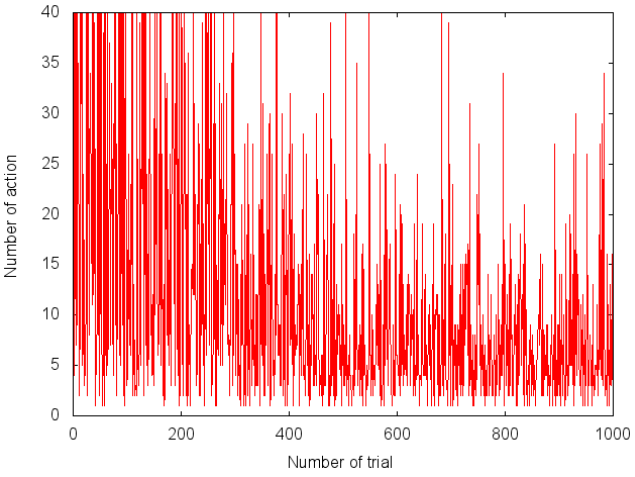


(c)：提案手法

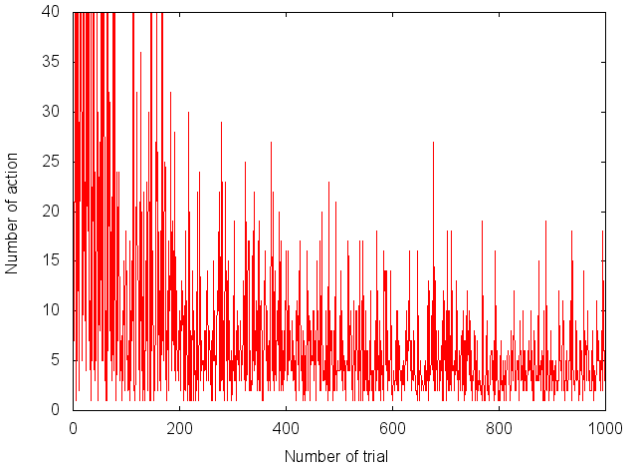
図 44：5 関節時の各手法の各試行での行動数の推移



(a)：シングルエージェント

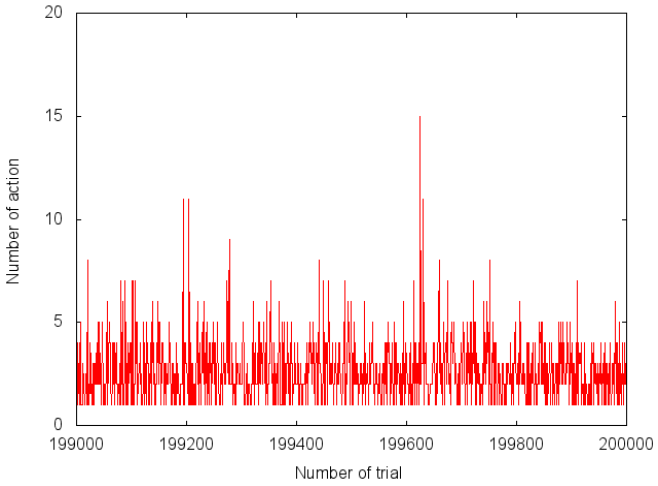


(b)：協調動作を学習しないマルチエージェント

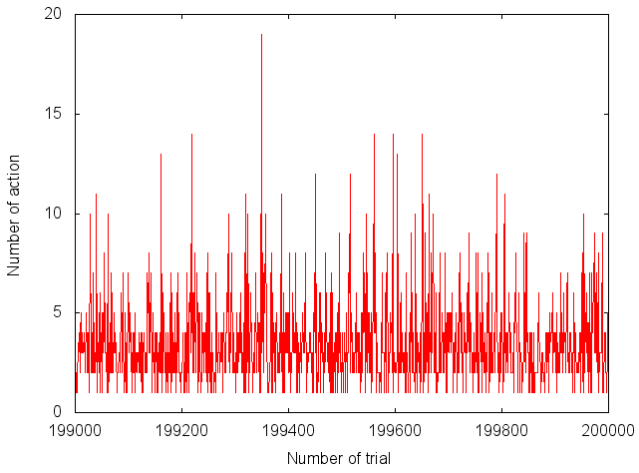


(c)：提案手法

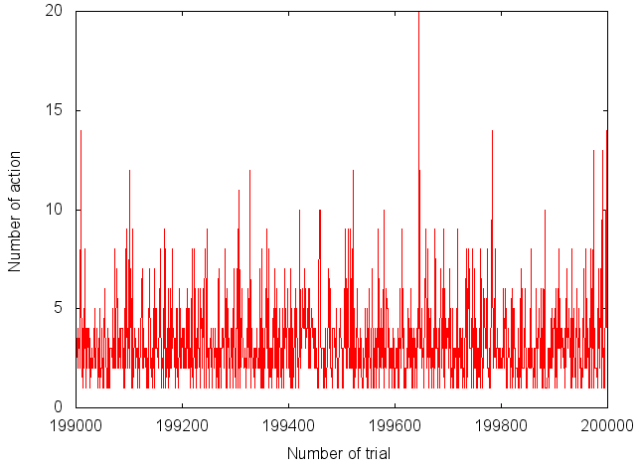
図 45：5 関節時の各手法の各試行での行動数の推移（1 試行から 1000 試行）



(a)：シングルエージェント



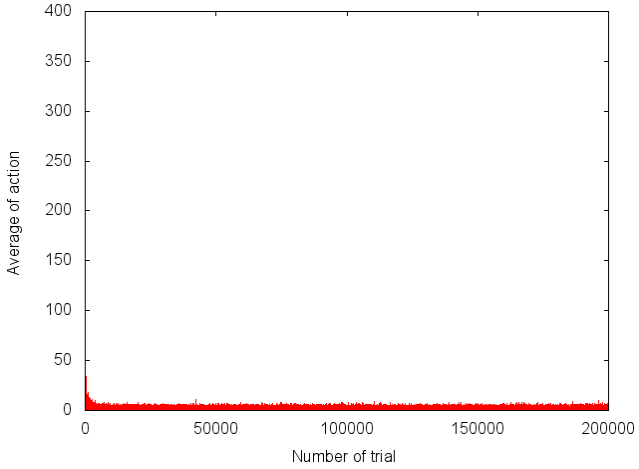
(b)：協調動作を学習しないマルチエージェント



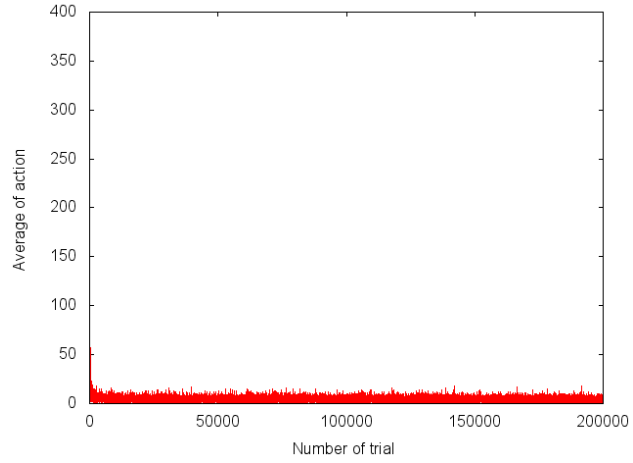
(c)：提案手法

図 46：5 関節時の各手法の各試行での行動数の推移（199000 試行から 200000 試行）

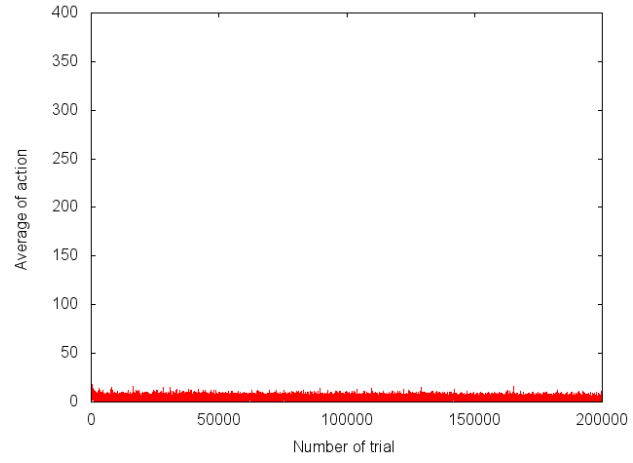
次に各手法の 3 試行の行動数の平均値の推移を図 47，図 48，図 49 に示す．横軸は試行数，縦軸は 3 試行の行動数の平均値となる．今回の実験では 1 回の行動で目標物体を回収することが可能な設定にしている．そのため 3 試行の行動数の平均値が 1 に収束していれば行動数が収束し学習が完了しているといえる．図 49 から 3 手法共に 3 試行の行動数の平均値が 1 付近で収束していることが分かる．したがって提案手法は従来手法 2 種と同等の行動を獲得していることが分かる．この結果から提案手法は試行を繰り返す中で学習を行い，シングルエージェントや協調動作を学習しないマルチエージェントと同等の行動を獲得することが示された．



(a)：シングルエージェント

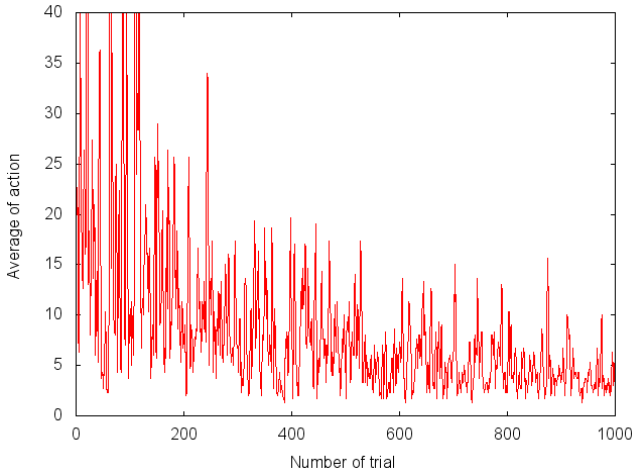


(b)：協調動作を学習しないマルチエージェント

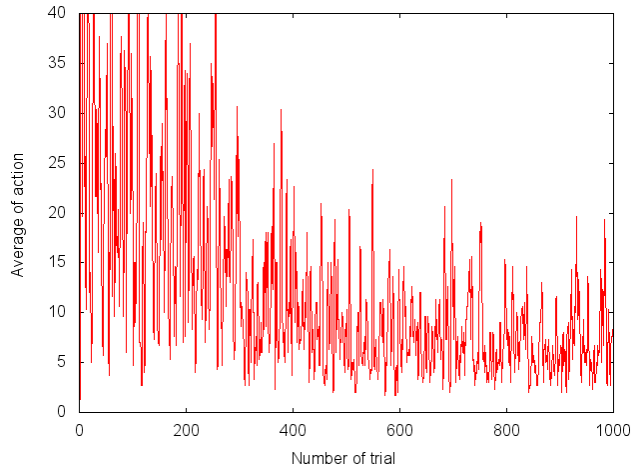


(c)：提案手法

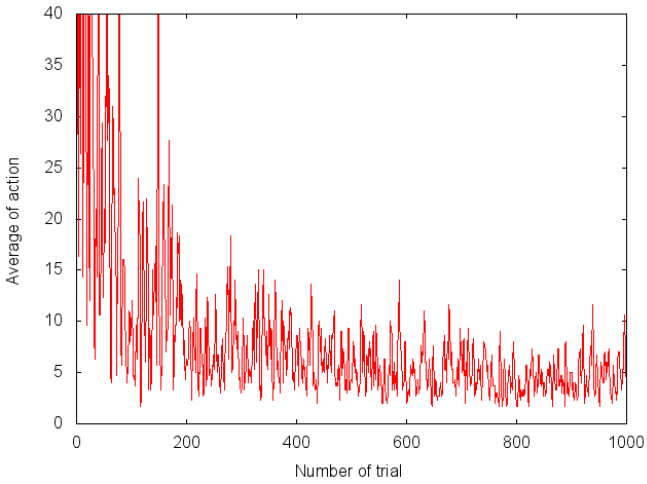
図 47：5 関節時の各手法の 3 試行での行動数の平均値の推移



(a)：シングルエージェント

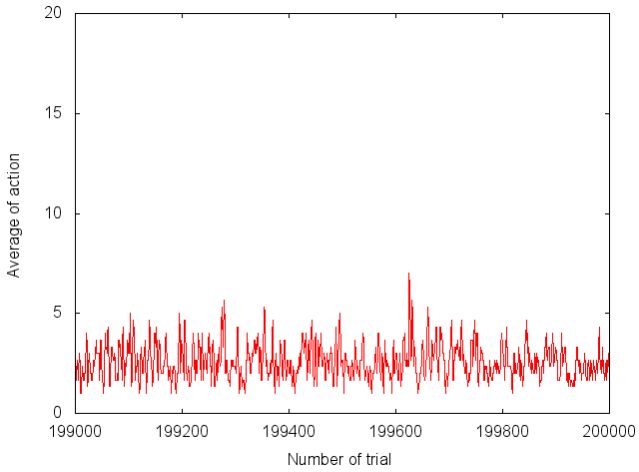


(b)：協調動作を学習しないマルチエージェント

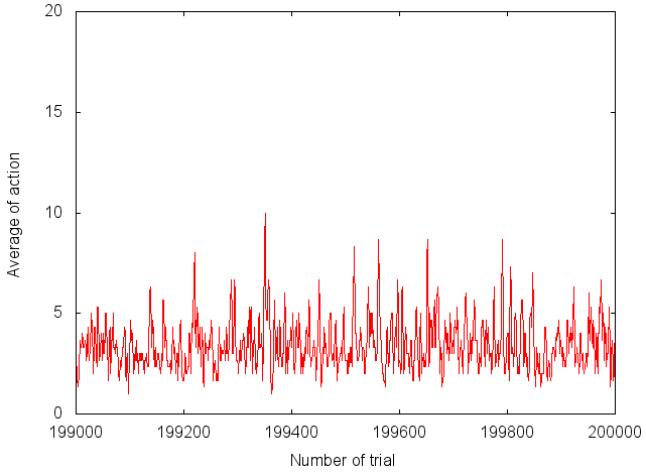


(c)：提案手法

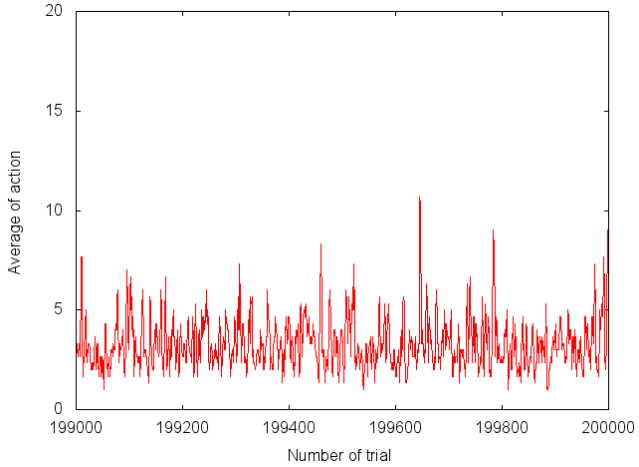
図 48：5 関節時の各手法の 3 試行での行動数の平均値の推移（1 試行から 1000 試行）



(a)：シングルエージェント



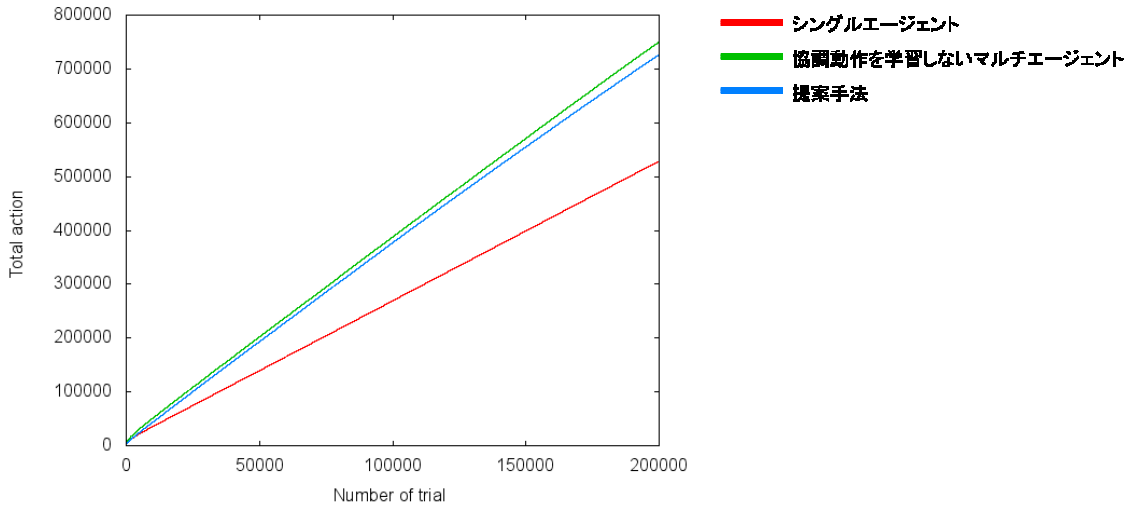
(b)：協調動作を学習しないマルチエージェント



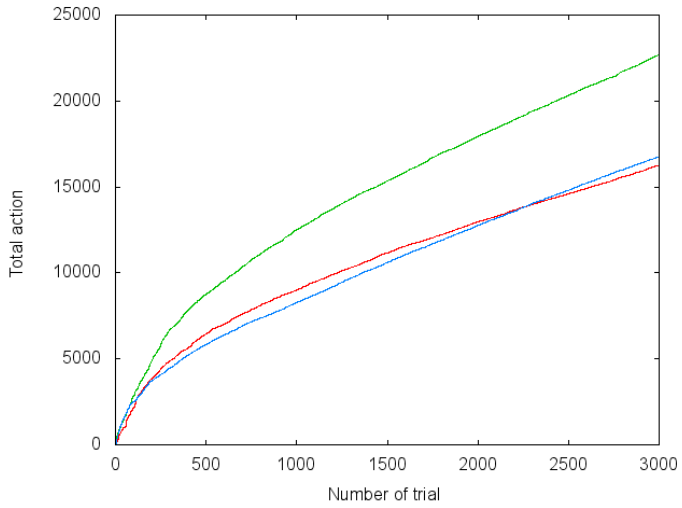
(c)：提案手法

図 49：5 関節時の各手法の 3 試行での行動数の平均値の推移（199000 試行から 200000 試行）

次に各手法の各試行時点での 1 試行目からの累計行動数を図 50 に示す．横軸は試行数，縦軸は各試行時点での累計行動数である．図 50 から提案手法の行動数の累計は協調動作を学習しないマルチエージェントと比較して少なくなっていることが分かる．このことから提案手法はエージェント間で協調動作を獲得してロボットの最適行動を選択できることが示された．一方でシングルエージェントの各試行時点での累計行動数と比較すると始めは提案手法の方が下回っていたが，2500 試行付近で累計行動数が逆転し最終的にはシングルエージェントが提案手法を下回るという結果になった．これは試行を繰り返す中でシングルエージェントが経験済みの状態行動対の割合が多くなり，最適行動を選択する回数が多くなったためと考えられる．



（a）：1 試行から 200000 試行



（b）：1 試行から 60000 試行

図 50：5 関節時の各手法の各試行時点での累計行動数の推移

次に各手法の各試行時点での経験済みの状態行動対の数と割合を図 51，図 52 に示す．横軸は試行数，縦軸は図 51 では各試行時点での経験済みの状態行動対の数，図 52 では各試行時点での経験済みの状態行動対の割合を示す．図 52 から協調動作を学習しないマルチエージェントは約 5 割以上の状態行動対を経験していることが分かる．一方でシングルエージェントと提案手法は 1 割にも満たない状態行動対しか経験していないことが分かる．この結果から協調動作を学習しないマルチエージェントは環境に対してある程度の状態行動対を経験しているが，シングルエージェントと提案手法に関しては十分な数の状態行動対を経験している状態にあるとはいえず，200000 試行のペースから考えてもこれ以上試行を繰り返したとしても十分な状態行動対を経験するのは難しいと考えられる．

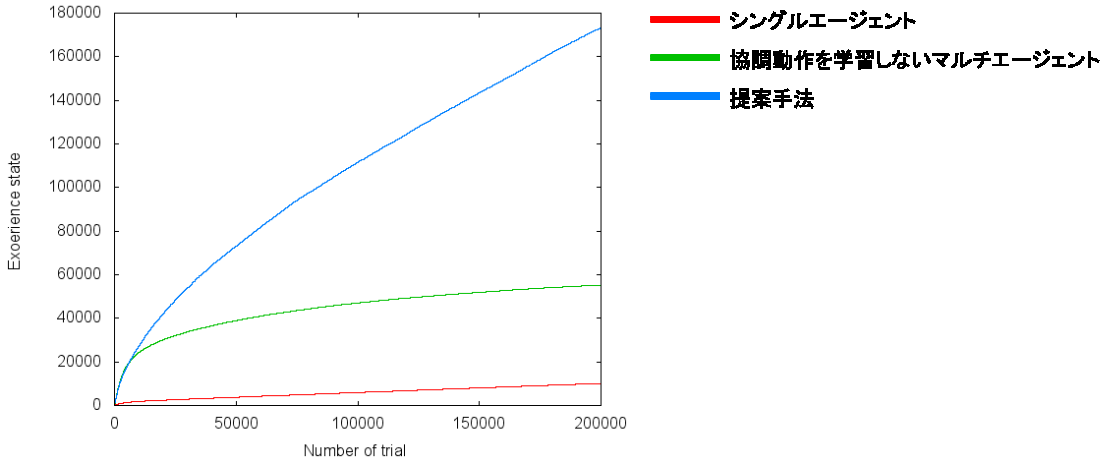


図 51：5 関節時の各手法の各試行時点での経験済み状態行動対の数

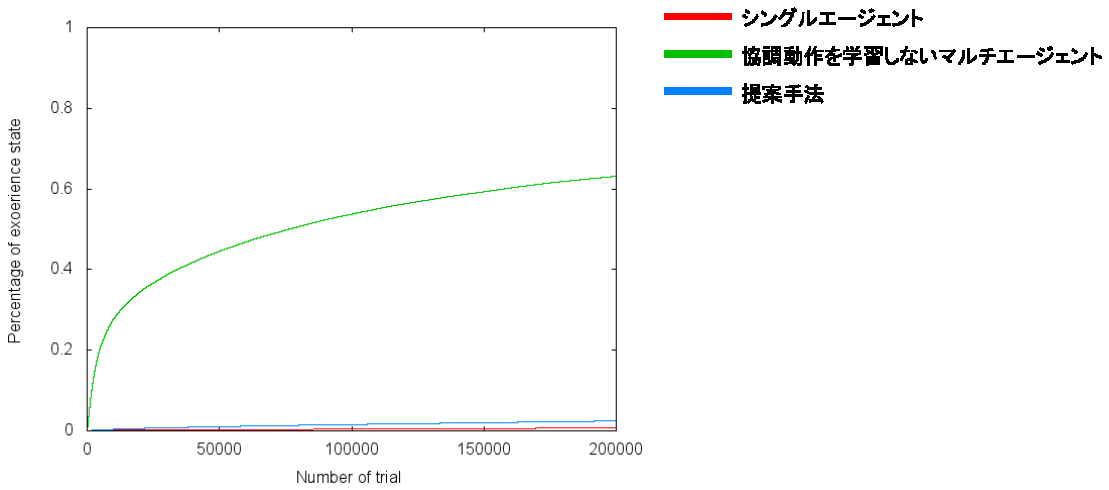


図 52：5 関節時の各手法の各試行時点での経験済み状態行動対の割合

4.4 考察

各関節における各手法の各試行の行動数の推移のグラフから，提案手法を用いたロボットは行動回数が収束し，タスク達成に必要となる行動を獲得していることが分かる．このことから提案手法は試行を繰り返すことでタスク達成に必要となる行動を学習することができることが示された．

また各試行時点での累計行動数の推移のグラフから，提案手法を用いたロボットは協調行動を学習しないマルチエージェントを用いたロボットと比較して，累計行動数が上回る場合と下回る場合に 2 種類が存在することが分かる．累計行動数が上回る場合は，ロボットアームの関節数が少なく最適行動が複数個存在する目標物体やある目標物体に対する最適行動の数が少ない場合である．一方で累計行動数が下回る場合は，ロボットアームの関節数が多く最適行動が複数個存在する目標物体の数やある目標物体に対する最適行動の数が多い場合である．このことから提案手法を用いたロボットの累計行動数が協調動作を学習しないマルチエージェントを下回るのは，ロボットが置かれる環境の状態数が多く，またタスク達成に必要となる最適行動が複数個存在する場合であるといえる．これは提案手法のエージェント間の協調行動の獲得によって協調行動を学習しないマルチエージェントではタスクを達成できない場合でもタスク達成に必要となる行動を獲得することができるということを示していると考えられる．したがって本研究の目標の 1 つである「各エージェント型のエージェントと協調した行動選択を学習するシステムを提案する.」また「この手法を単体のロボットに対して適用することで，最適行動が複数個存在するタスクをロボットが実行する場合，複数の最適行動の中から各エージェントが協調することで 1 つの行動を選択すること.」が達成することができたと考えられる．

一方で，2 関節のロボットアームの場合で各手法の各試行時点での経験済みの状態行動対の割合から，3 手法共に環境に対して十分な状態行動対を経験しているにもかかわらず，提案手法の各試行時点での累計行動数が従来手法 2 を上回ることが示された．これは今回提案手法を用いたロボットで採用した探査的行動の選択方法が関係している．提案手法ではロボット 1 回の行動選択に対して各エージェントがεの確率でランダム行動を選択している．そのため実際にロボットの選択した行動の中にランダム行動によって選択されたアクチュエータの動作が含まれる確率はεの値より高くなることが分かる.そのため提案手法とシングルエージェントではεの値が等しい場合,提案手法の方が選択された行動にランダム要素が含まれる確率が高くなるのである．また 2 関節のロボットアームの場合では環境全体の状態数の数も少なく，また最適行動が複数個存在する目標物体の数やある目標物体に対する最適行動の数も少ない．そのためエージェント間の協調行動を獲得しなくてもタスクを達成するのに必要となる行動を選択することが簡単である．したがって提案手法と協調動作を学習しないマルチエージェントを比較した場合，各エージェントの状態行動対の数が少なくより少ない試行数で多くの状態行動対を経験することができる協調動作を学習

しないマルチエージェントの方が，累計行動数が少なくなるのである．

さらに関節数 4 つと 5 つのロボットアームの各手法の各試行時点での経験済みの状態行動対のグラフから，シングルエージェントと提案手法の経験済みの状態行動対の割合が環境に対して十分でないということが示された．これはロボットアームの関節数が多くなると環境内の状態行動対の数も多くなり，シングルエージェントと提案手法では環境に対して十分な状態行動対を経験することが困難であると考えられる。今回使用した実験タスクの設定によって環境に対して十分な状態行動対を経験するのが難しいということも理由の 1 つに挙げられる．一方で協調動作を学習しないマルチエージェントの場合では各試行時点での経験済みの状態行動対の割合が他の手法 2 種と比較して高い割合を出していることが分かる．このことから提案手法は協調動作を学習しないマルチエージェントと比較して環境に対して十分な状態行動対を経験するまでの試行数が増加する，または十分に経験することが困難となることが示された．この理由として，1 つはエージェント間の協調行動を学習するために各エージェントの状態に他のエージェントの選択した行動を加えたことが上げられる．状態が増えたことで各エージェントの状態行動対が増加し，その結果各エージェントが環境に対して十分な経験を積むまでの試行数が増加するのである．

以上の結果から，提案手法が学習によってタスク達成に必要となる行動を獲得することが示された．また提案手法はエージェント間の協調行動を学習することができる．そのため協調行動を学習しないマルチエージェントと比較して環境に対して十分な経験を積んだ状態である場合，ロボットの最適行動を一意に選択できる．特に環境状態の数が多く，またタスク達成に必要な行動が複数存在する場合提案手法が有効に働く．一方で提案手法はエージェント間の協調行動を学習するために各エージェントの状態に他のエージェントの行動を加えた．その結果各エージェントの持つ状態行動対が増加したため，協調行動を学習しないマルチエージェントと比較して経験済みの状態行動対の割合が下がってしまう．ただし提案手法の各エージェントの持つ状態行動対の数は，シングルエージェントの場合の 1 エージェントが持つ状態行動対の数より多くなることはない．そのためシングルエージェントよりも経験済みの状態行動対の割合が極端に少なくなることはないと考えられる．

第 5 章　まとめ

5.1　論文全体の考察

本研究では，先行研究の問題点からエージェント間の協調行動による単体ロボットの行動獲得を行うシステムの構築を目標とした．エージェント間の協調行動の獲得する方法の 1 つとして，本研究では各エージェントが選択した行動に注目した．各エージェントの所持する状態行動対の状態に，他のエージェントが選択した行動を加えることで，各エージェントが他のエージェントが選択した行動に対して行動を選択できるようにした．そこで本研究は各エージェントが所持する状態に他のエージェントが選択する行動を加えることでエージェント間の協調動作を獲得するアプローチをとった．各エージェントが，他のエージェントが選択した行動を取得するためには 1 度各エージェントが行動選択を行う必要がある．そこで本研究ではロボットの 1 回の行動選択時にロボット内の各エージェントが複数回行動選択を行う方法を考えた．エージェントの行動選択時に各エージェントは選択した行動を他のエージェントに送信する．各エージェントが他のエージェントが選択した行動を受け取ることで他のエージェントが選択した行動に合わせて行動を選択することができる．そこで本研究ではロボットの 1 回の行動選択時に各エージェントが複数回行動選択を行い，エージェントが選択した行動を共有することでエージェント間の協調動作を獲得するロボットの行動獲得手法を提案した．

本研究ではエージェント間で選択した行動を送信して協調行動を獲得するために，エージェント間の情報共有とエージェントの行動選択の一連の流れをステップと定義し，ロボットの 1 回の行動選択時にはこのステップを既定の回数繰り返すことで他のエージェントが選択した行動の取得を可能とした．このステップを定義することで，エージェントは複数回他のエージェントが選択した行動を元に行動選択を行うことができる．そのため 1 回の行動選択ではエージェント間で選択した行動がロボットの最適行動に揃わない場合でも，何度も行動選択を行うことでエージェントの行動が一意に定まるまで行動選択を行うことができる．また各エージェントには強化学習によって定義される行動評価値とは別に，行動遷移確率を定義した．行動遷移確率とは各エージェントの行動評価値を元に算出したものであり，各エージェントの最適行動に対する重みを表す．この行動遷移確率は各ステップで更新され，各エージェントが各ステップで選択した行動に対して他のエージェントが選択した行動と選択しなかった行動に対してそれぞれ重みを加えていく．各エージェントは行動評価値と行動遷移確率を元に行動を選択する．行動遷移確率を定義行動選択時に利用することで他のエージェントが 0 ステップ目から直前のステップまでに選択した行動の傾向を蓄積することができる．そのため各エージェントは他のエージェントが選択する行動の傾向から行動を選択することができるようになり，行動を一意に定めることができるようになる．

本研究では提案手法による行動選択によってロボットがタスクを達成できることと，本研究の目的であるエージェント間の協調行動の獲得が達成できたかの確認を目的にシミュレーション実験を行った．実験ではロボットアームタスクのリーチング動作による目標物体回収タスクを行った．ロボットアームには関節を稼働させるサーボモータが搭載されており，各関節を稼働することでロボットアームを動かすことができる．ロボットアームの先端には物体を回収できるハンドが搭載されている．このハンドで目標物体の位置に合わせることがタスクの目的となっている．本実験のタスクでは目標物体の発生位置によっては最適行動が複数存在する場合がある．そのためロボットアームにマルチエージェントシステムを適用した場合，複数の最適行動の中から一意に行動を決定しなければならない．そのためエージェント間の協調行動を獲得できなければタスク達成が困難なものとなっている．

実験結果から，提案手法を用いたロボットアームは行動回数が収束し，従来手法を用いたロボットアームと同等の行動を獲得することが示された．また提案手法を用いたロボットアームは協調動作を学習しないマルチエージェントと比較して累計行動数が少なくなる場合があることが示された．その累計行動が少なくなる場合は状態数が多く最適行動の数が複数存在する時である．このことから提案手法はエージェント間の協調行動を獲得することができたと考えられる．以上の結果から提案手法が強化学習によってタスク達成に必要となる行動を獲得でき，エージェント間の協調行動を獲得しロボットの最適行動を選択するという目標を達成した．

5.2 今後の課題

本節では本研究で提案した手法の今後の課題について説明する．

5.2.1 他の機械学習への適用

本研究で提案した学習手法は強化学習であり，その中でも$Q-learning$と$\varepsilon-greedy$法を用いて提案手法を構築し実験を行った．そのため提案手法が使用可能である事が確認されているのは強化学習の中でも$Q-learning$法と$\varepsilon-greedy$法を使用した強化学習のみである．しかし，単体ロボットに対するマルチエージェントシステムという枠組みは他の強化学習手法，さらには他の機械学習手法に適用できる可能性がある．また強化学習内でもほかの行動選択手法や行動学習手法を適用できる可能性もある．特に行動選択手法については探査的行動を選択する範囲についても検討することができる．それによって同じ学習手法であってもロボット全体の挙動が変化する可能性がある．提案手法が他の機械学習や学習手法にも適用可能であることが証明されたならば，機械学習全般で使用可能な手法となることが期待できる．またタスクやロボットの種類に応じて適切な学習手法を選択できればよりタスクに対応したロボットを開発することができる．

5.2.2 実ロボットへの適用

強化学習は実ロボットへの適用に適している手法である．本研究で行った実験も実ロボットを仮定したシミュレーション実験であった．そのため提案手法も実ロボットへの適用が期待できる．しかし実ロボットによる実験，検証を行っていないため，実ロボットに適用した際の行動学習は保証されていない．また実ロボットに適用，シミュレーションでは発生しない実ロボット特有の問題も発生する可能性がある [7]．特に本研究で提案した手法はエージェント間の協調動作を獲得するためにエージェント間で通信を行う必要が存在する．この時エージェント間の通信が行動学習にどのように影響を与えるかを検証する必要がある．これらの要素から実ロボットに適用し実験を行い，実ロボットに適用した際の問題点を発見し，それを解決することが必要となる．

5.2.3 未知の状態行動対の経験

実験結果から提案手法は協調行動を学習しないマルチエージェントと比較して各試行時点での経験済みの状態行動対の割合が少ないことが示された．その理由として提案手法では各エージェントの状態に他のエージェントが選択した行動を加えているため状態の数が増加している点があげられる．各試行時点での経験済みの状態行動対の割合が少ないとロボットがタスク達成に必要となる行動を学習するまでの試行数が増加するという問題点につながる．

この問題点を解決する方法の 1 つとして，始めに協調動作を学習しないマルチエージェントで学習を行い，ある程度状態行動対を経験した後に提案手法を用いてエージェント間の協調動作を学習する方法を提案する．提案手法と協調行動を学習しないマルチエージェントでは，同一環境，同一ロボットでの同一タスクという条件下ならば，他のエージェントが選択した行動以外の環境状態と出力する行動が同じである．そのため他のエージェントが選択した行動の移行の仕方のみ定義することでエージェントの行動評価値の移行は可能である．そのため協調動作を学習しないマルチエージェントで学習した行動評価値を提案手法で利用することが可能と考えられる．この方法を用いることで協調動作を学習しないマルチエージェントの状態行動対の経験率の高さと提案手法のエージェント間の協調動作の獲得を両方利用した手法を定義することが可能である．

5.2.4 各エージェントの探査的行動選択の割合

2 関節ロボットアームの実験結果から提案手法を用いたロボットは環境に対して十分な状態行動対を経験した状態ではあるが，従来手法 2 種と比較して各試行時点での累計行動数が多くなっているという結果が示された．この理由として提案手法では各エージェントがεの確率でランダムに行動選択を行うという設定にしている点があげられる．この設定の場合ロボットの行動の中にランダム要素によって選択された行動が含まれる確率は設定したεの値よりも高くなる．そのためシングルエージェントと比較して最適な行動を経験

済みの状態にある時にその最適行動を選択できる確率が下がってしまうという問題点が発生する.

この問題点を解決する方法としては以下の方法が考えられる

1. 探査的行動選択の割合を設定するエージェントの数に応じて決定する.
2. 探査的行動選択かどうか決定する範囲を変更する.

1 の方法は, ロボットに設定するエージェントの数に応じて探査的行動選択を選択する割合を決定する方法である. 例えば $\varepsilon - greedy$ 法であれば ε の値を式(6)のように決定する. このように決定することで設定するエージェントの数に合わせた ε の値を設定できる.

$$\varepsilon = \varepsilon' / (\text{エージェント数}) \qquad \cdots (6)$$

2 の方法は, ロボットの行動選択が最適な行動か探査的行動かを選択する範囲を変更し, タスクにあった範囲を設定するという方法である. 本研究で設定した範囲はロボットの 1 回の行動選択時に各エージェントが選択するものであった. しかしそれ以外の範囲として以下の範囲が考えられる.

1. ロボット 1 回の行動選択時にロボット全体が最適行動か探査的行動か選択.
2. 各ステップの行動選択時に各エージェントが最適行動か探査的行動か選択.

探査的行動選択かどうかの範囲が変わることで, ロボットが探査的行動を選択する確率も変化する. ロボットの種類やタスク, 設定するエージェントに応じて最適な範囲を設定することでより状況に合わせたロボットを開発することができるようになる.

参考文献

[1] 米本完二, “産業用ロボットの今後の技術動向,” 日本ロボット学会誌, pp7-13, 1993-1-15.

[2] 水川真, 小山俊彦, “産業用ロボットの教示方法の現状と展望,” 日本ロボット学会誌, pp180-185, 1999-3-15.

[3] 藤田雅博, “Robot Entertainment System AIBO の開発,” 情報処理, pp146-150, 2000-02-15.

[4] B. Richard S.Sutton, Andrew G, 強化学習, 森北出版, 2000-12.

[5] 畝見達夫, “強化学習,” 人工知能学会誌, Vol.9, No.6, pp830-836, 1994.

[6] 畝見達夫, “強化学習法とロボットへの応用,” 日本ロボット学会誌, Vol.13, No1, pp51-56, 1995.

[7] 浅田稔, “強化学習の実ロボットへの応用とその課題,” 人工知能学会誌, 12(6), 831-836, 1997-11-01.

[8] 浅間一, “マルチエージェントロボットシステム研究の動向と展望,” 日本ロボット学会誌, Vol.10, No.4, pp.428~432 1992.

[9] 淺間一, “マルチエージェントから構成された自律分散型ロボットシステムとその協調的活動,” 精密工学会誌, 57(12), pp2117-2122, 1991.

[10] 三上貞芳, “強化学習のマルチエージェント系への応用,” 人工知能学会誌, 12(6), pp845-849, 1997.

[11] 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, 13(4), pp609-618, 1998.

[12] 荒井幸代, “マルチエージェント強化学習 : 実用化に向けての課題・理論・諸技術との融合（＜特集＞「マルチエージェント技術における新しい可能性」）,” 人工知能学会誌, 16(4), pp476-481, 2001.

[13] 参沢匡将, 木村春彦, 廣瀬貞樹, 大里延康, “強化学習型マルチエージェントによる交通信号制御,” 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, pp478-486, 2000.

[14] 森紘一郎, 山名早人, “強化学習並列化による学習の高速化,” 情報処理学会研究報告. ICS, pp151-155, 1991.

[15] 高泉昇太郎, 倉重健太郎, “マルチエージェント強化学習によるシングルロボットの行動学習,” 日本ロボット学会第 30 回記念学術講演会, RSJ2012AC4F1-6, 札幌, 北

海道, 2012.9.17-20.

[16] 伊藤昭，金渕満，“知覚情報の素視化によるマルチエージェント強化学習の高速化-ハンターゲームを例に-,” 電子情報通信学会氏論文, No.3，pp.285-293，2001.

[17] 岡本昌紘，杉山久佳，辻岡哲夫，村田正，“災害救助を目的とした低電力消費型群ロボットネットワークシステム,” 電子情報通信学会技術研究報告．CS，通信方式, 105(280)，pp31-36，2005.

[18] 勝本悟史，五十嵐洋，“複数台移動ロボットによる協調捕獲,” ロボティクス・メカトロニクス講演会講演概要集, 2007.

[19] 森啓，納谷太，大里延康，“6 軸マニピュレータの分散制御実験,” 電子情報通信学会総合大会講演論文集, 167，1996.

[20] 小鍛冶繁，“多自由度機構と分散制御,” 精密工学会誌, 54(10)，pp1921-1926，1988.

[21] 佐藤貴英，加納剛史，石黒章夫，“局所的な齟齬情報に基づくヘビ型ロボットの適応的自律分散制御方策,”ロボティクス･メカトロニクス講演会講演概要集, 1A1-F02(1), 2010.

[22] 光本直樹，福田敏男，荒井史人，“マルチエージェントシステムにおける群戦略の生成・適正メカニズムに関する研究：免疫クローン選択形群ロボットによる群戦略の最適化,” 日本機械学會論文集. C 編, 61(586)，pp-2440-2447，1995.

[23] 本多勝明，村上国男，“交通信号系の分散協調制御,” 電子情報通信学会秋季大会講演論文集, 141，1994.

[24] 柴田克成，杉坂政典，伊藤宏司，“強化学習によるリーチング動作の獲得,” 電子情報通信学会技術研究報告．NC，ニューロコンピューティング, 100(688)，pp107-114，2001.

謝辞

本論文を結ぶにあたり，日ごろより懇切なるご指導を賜りました倉重健太郎先生に深く感謝の意を表します．また，ご助言，ご指導をいただいた畑中雅彦先生，佐賀聡人先生，本田泰先生に感謝の意を表します．そして論文の査読や助言をしていただいた認知ロボティクス研究室の木島康隆さん，杉本大志さん，中南義典さん，宮崎愛央さん，北山直樹さん，渋谷和さん，梅津祐介さん，沼田利伸君，三浦丈典君，木村敏久君，挾間重直君，平間経太君，二階堂芳君，片山和宣君，小橋遼君，千葉秀平君に感謝します．

研究業績

[1] 高泉昇太郎，倉重健太郎，“マルチエージェント強化学習によるシングルロボットの行動学習”，日本ロボット学会第 30 回記念学術講演会，RSJ2012AC4F1-6，札幌，北海道，2012.9.17-20