目次

第１章　序論

本章では研究背景として，ロボットの歴史や強化学習，マルチエージェントシステムについて述べる．そして，従来研究と先行研究を説明し本研究の目的を述べる．

1.1　ロボットの歴史

ロボットが現実社会で用いられるようになったのは 1960 年ごろである．この頃のロボットは主に人間の代わりに単純な繰り返し作業や人間では作業しにくい悪環境下での苦渋作業を行うことに使用されていた．このロボットを産業用ロボットと言う．産業用ロボットは第一世代のロボットに分類される．産業用ロボットは工場などで用いられ，作業の効率化に貢献した．しかし，産業用ロボットは複雑な作業を行うことはできなかった．

産業用ロボットが登場してからもロボットの研究は盛んに行われていった．そして，人間の五感に相当するセンサを装備することで複雑な行動をとれるようなロボットが登場した．このロボットは第二世代のロボットに分類される．センサは周辺の情報などを感覚情報として信号に変換する．このセンサをロボットに搭載することで，センサから得られた信号に応じて環境に対応した動作を行うことができるようになった．これによりロボットの使用できる場面が増えた．

さらに研究が進むと，ロボットが自身の経験から自身の行動を制御，修正できるロボットが登場した．このロボットを学習制御ロボットと言い，第三世代のロボットに分類される．学習制御ロボットの研究が進めば，事前に人間による設定を必要とせず，ロボットが何らかの情報を得ることにより，自身の行動制御方法を自律的に獲得可能になることが期待されている．学習制御ロボットが実用化される段階になれば，ロボットの活躍する場面も更に拡大することが期待されている．学習により行動を制御できるようになることで，人間の生活空間のような多様な環境の中で複雑で柔軟性のある振る舞いをロボットがとることができる．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扱っている．

1.2　機械学習

機械学習[1]は，およそ 1960 年あたりから人工知能の分野の 1 つとして研究が始まった．人工知能研究の一環として，機械が試行錯誤により学習することで自律的に行動を獲得していく仕組みを実現することを目指す分野である．現在では，数値・文字・画像・音声など多種多様なデータの中から，規則性・パターン・知識を発見し，現状を把握や

将来の予測を行うために発見した知識を役立てることが目的となっている. 機械学習は大きく分けて，教師あり学習，教師無し学習，強化学習の 3 つに分類される．

教師あり学習は学習データを用いて学習する手法である. 教師あり学習では入力データに対する望ましい出力データが既知である．教師あり学習の目的は，環境から得た学習データの入出力関係から，学習背景にあるその環境を学習することである．

一方で教師なし学習は学習データを用いて学習するが, 教師なし学習では入力データに対する望ましい出力データが未知である. 与えられた未知のデータから特徴的なパターンを見つけ出し，データの本質的な構造を抽出するために用いられる．

強化学習[2][3]はある環境内におけるエージェントが，現在の状態を観測し，取るべき行動を決定する問題を扱う機械学習の一種である. 強化学習で扱うエージェントは学習を行う主体のことである．強化学習は，教師あり学習と異なり学習のための適切な入力データと出力データのペアが与えられることがない. 強化学習では行動に応じて環境から報酬を得ることで学習を進めていく．報酬はスカラー値で与えられ，エージェントは報酬を出来るだけ多く得られる行動を学習する. 強化学習は未知の学習領域を開拓していく行動と, 既知の学習領域を利用していく行動とをバランス良く選択することができるという特徴を持っている. その性質から未知の環境下でのロボットの行動獲得に良く用いられる[4]．このことから本研究は機械学習のなかでも強化学習を扱っている．強化学習についての詳しい説明は第 2 章で行う．

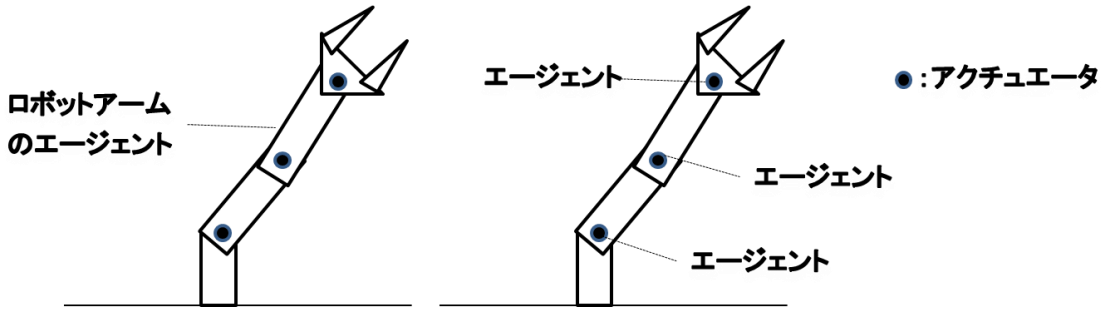
1.3 マルチエージェントシステム

マルチエージェントシステム[5][6]とは，複数のエージェントで構成されたシステムのことである．ここでのエージェントは制御を行う主体のことを言う．近年の計算機アーキテクチャや情報処理手法, ハードウェアの急激な進歩に支えられた並列分散処理技術の発展を背景に生まれてきた考え方である. 複数のエージェントを並列分散的に活動させ，単一の場合では不可能な，あるいは単一の場合よりも質の高い機能を実現しようとする手法である．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ある．そのため，システム全体の振る舞いは各エージェントの意思に大きな影響を受ける. 各エージェントが他のエージェントを考えずに振る舞うとシステム全体としての目標を達成できなくなる. 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振る舞いを考慮し, 他のエージェントの振る舞いに合わせた意思決定を行うことが重要となる. 本研究ではマルチエージェントシステムに強化学習を適用する手法について扱う．マルチエージェントシステムについては 3 章で詳細を述べる．

1.4　従来研究と先行研究

マルチエージェントシステムはロボットの制御にも用いられる．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がのうち，1 つのロボットにマルチエージェントを設定した研究[7-8]が存在する．これらの研究では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．

マルチエージェントシステムに強化学習を適用した先行研究として，1 つのロボットに複数の強化学習エージェントを設定した研究がある[9]．先行研究の手法の概要を図 1 に示す．先行研究の手法はロボットに搭載されているアクチュエータに一対一でエージェントを設定する．ロボットに搭載されているアクチュエータは複数あるので，エージェントは複数存在することになる．各エージェントは設定されたアクチュエータの動作を学習する．エージェントをアクチュエータ毎に設定することで，ロボットの行動を複数のエージェントで分担する．これによってエージェント 1 つでロボットの行動を学習する場合と比較して，マルチエージェントでは 1 つのエージェントが所持する行動数が削減される．その結果エージェント 1 つ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できる．



（a）従来の強化学習　　（b）先行研究

図 1.先行研究のマルチエージェントシステム

先行研究でのマルチエージェントシステムの学習の流れを説明する．まずロボットは現在の環境状態を観測する．次にロボットは環境状態を各エージェントに送る．各エージェントは受け取った環境状態を基に自身に割り当てられたアクチュエータの行動を選択する．全エージェントが行動を選択した後，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一斉に出力する．各エージェントは環境から報酬を受け取る．各エージェントは受け取った報酬を基に自身が選択した行動について学習する．この流れを繰り返していき各エージェントは割り当てられたアクチュエータの動作を学習する．

1.5 先行研究の問題点

先行研究の問題点として, 従来の強化学習と比べて学習が進んでからの行動数が全体的に多くなることが挙げられる. この問題の起こる原因として, 先行研究では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. 先行研究の手法では, 各エージェントはそれぞれ独立して行動選択を行っている. 報酬が得られるかどうかはロボット全体での行動によって決まる. そのため, ロボットの最適行動が複数存在する場合には各エージェントがそれぞれ自身にとっての最適行動をとったとしても, ロボットの行動が最適行動とならない場合がある. したがって各エージェントが目的達成に必要な行動を学習済みの状態であっても, 各エージェントが最適行動を取ろうとした時にロボットの行動が一意に定まらないことが起こりえる.

この問題を解決するには各エージェントが行動を選択するときに環境状態だけでなく他のエージェントの行動の情報も考慮する必要がある. 各エージェントが他エージェントの行動を考慮に入れて協調することで, ロボットの最適行動が複数存在するタスクでも, ロボットの複数ある最適行動の中から特定の 1 つの行動になるように自身の行動を選択できる.

1.6 研究目的

本研究ではシングルロボットにマルチエージェントシステムを用いた強化学習を適用し, 各エージェントが他のエージェントと協調して行動選択をするシステムを提案する. この手法を適用することで, 最適行動が複数存在するタスクをロボットが学習する場合に, 各エージェントが協調して行動選択をすることでロボットの行動を一意に決めることができる. これにより, 先行研究の問題点である学習が収束してからの行動数のばらつきが多いことを解決する.

1.7 提案手法のアプローチ

マルチエージェントシステムの重要な点として各エージェントが協調して行動する点が挙げられる. マルチエージェント環境では各エージェントが協調しなければ, 他エージェントの妨害をするなどして全体としての目的を達成することが困難となるからである. そのためマルチエージェントシステムでは, 各エージェントが協調する動作を組み込むことが必須となる. 各エージェントで協調的行動を行うためには, エージェント間で情報のやり取りを行い, その情報を行動選択に利用する必要がある.

本研究で提案するマルチエージェントシステムでは, 各エージェントの協調する手法として, 各エージェントの環境状態に他エージェントの行動の情報を加えることを考え

た. そうすることで各エージェントは他エージェントが選択した行動を状態の 1 つとして認識する. 環境状態に他エージェントの行動を加えることで, 各エージェントは他エージェントの行動を含めて, 自身の行動を評価することが可能となる.

しかし, ただ環境状態に他エージェントの行動の情報を加えて強化学習を適用するだけでは不十分である. 理由として環境状態を取得する時には他エージェントがどのような行動をとるかはわからない点が挙げられる. 他エージェントの選択する行動を知るためには少なくとも一回は全エージェントが行動を選択している必要がある.

そこで本研究ではロボットの一回の行動選択の際に各エージェントが複数回仮想的に行動選択を行う手法を考える. 各エージェントが複数回仮想的に行動選択し, その都度他エージェントと選択した行動の情報の送受信を行う. これにより, 他エージェントの行動の情報を考慮して行動選択できるようになる. 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の情報を通信していく動作を本研究では反復合議と呼ぶ. 手法についての詳しい説明は 4 章にて述べる.

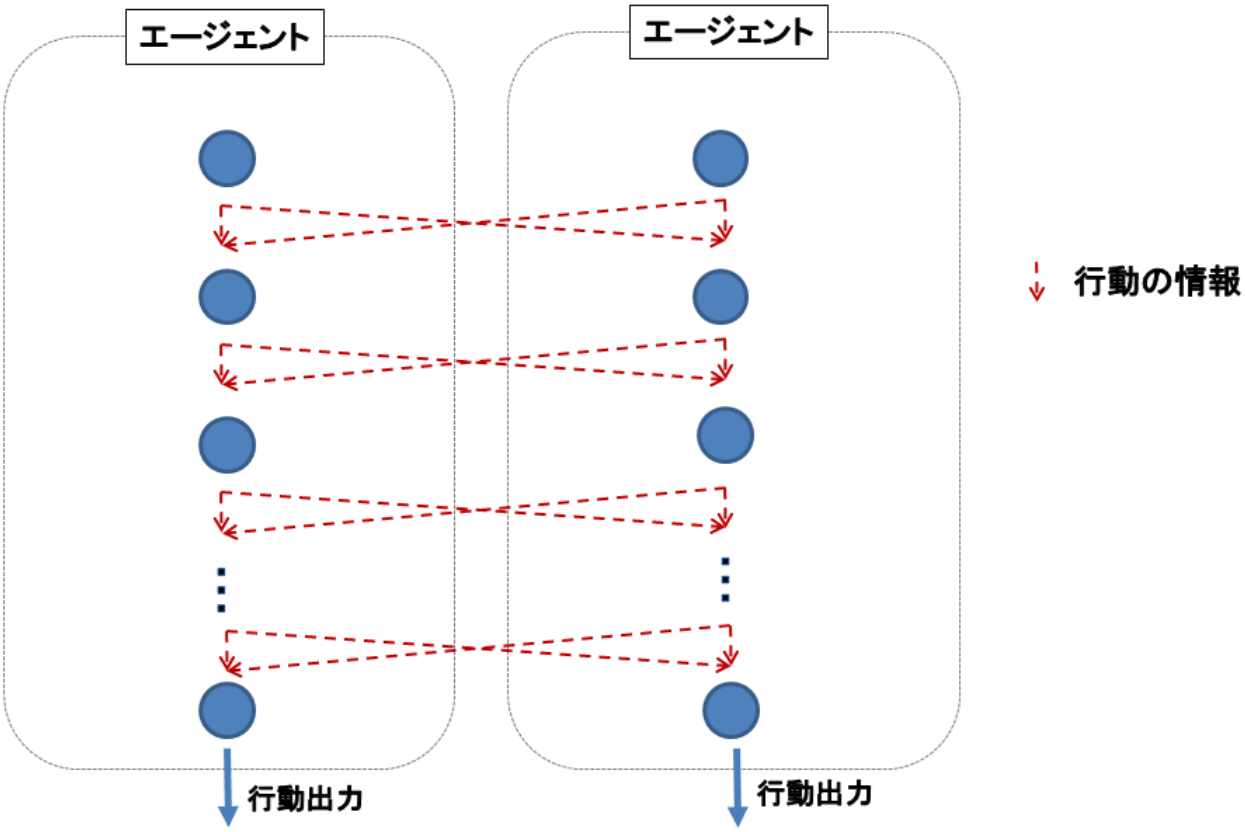


図 2.提案手法のアプローチ

1.8 論文構成

第 1 章では序論として，研究背景にあるロボットの歴史，機械学習，マルチエージェントシステムについて説明した. またマルチエージェントを用いた強化学習のロボットの適用例として先行研究の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た．そして先行研究の問題点から本研究の目的を述べた.

第 2 章では強化学習の概要と具体的な手法について説明する. 強化学習に関わる行動選択手法についてもこの章で説明する.

第 3 章ではマルチエージェントシステムの概要とマルチエージェントシステムを用いた強化学習について説明する．最後に本研究のアプローチを述べる.

第 4 章では本研究で提案するマルチエージェントシステムを用いた強化学習をシングルロボットに適用する手法について説明する．本章ではシステムの概要，システムの流れ，協調手法の説明をする.

第 5 章では提案手法の検証をするための実験について説明する．本章では実験目的，実験概要そして実験設定を説明する. その後実験結果を示し, 実験結果の考察を述べる. 実験ではロボットアームによるリーチングタスクのシミュレーションを行った.

第 6 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も述べる.

第２章　強化学習

本章では，強化学習の概要と強化学習に関わる行動選択手法について説明する．また強化学習の 2 つの分類である環境同定型と経験強化型についても説明する．特に環境同定型の強化学習としては本研究で用いている Q 学習のアルゴリズムについて説明する．

2.1　強化学習の概要

強化学習は，ある環境内におけるエージェントが，現在の状態を観測し，取るべき行動を決定する問題を扱う機械学習の一種である．強化学習の概要を図 3 に示す．エージェントは学習を行う主体のことである．強化学習における学習の流れを図 4 に示す．エージェントは環境状態を観測する．観測した状態を基に行動を選択する．行動に応じてエージェントは環境から報酬を得る．報酬は目的を達成することで与えられる．報酬はスカラー値で与えられる．エージェントは得られた報酬を基に自身の行動評価値を更新し，学習を進めていく．行動評価値はある環境状態におけるエージェントの行動の評価値である．強化学習ではこの一連の動作を通じて報酬が最も多く得られるような方策を学習する．

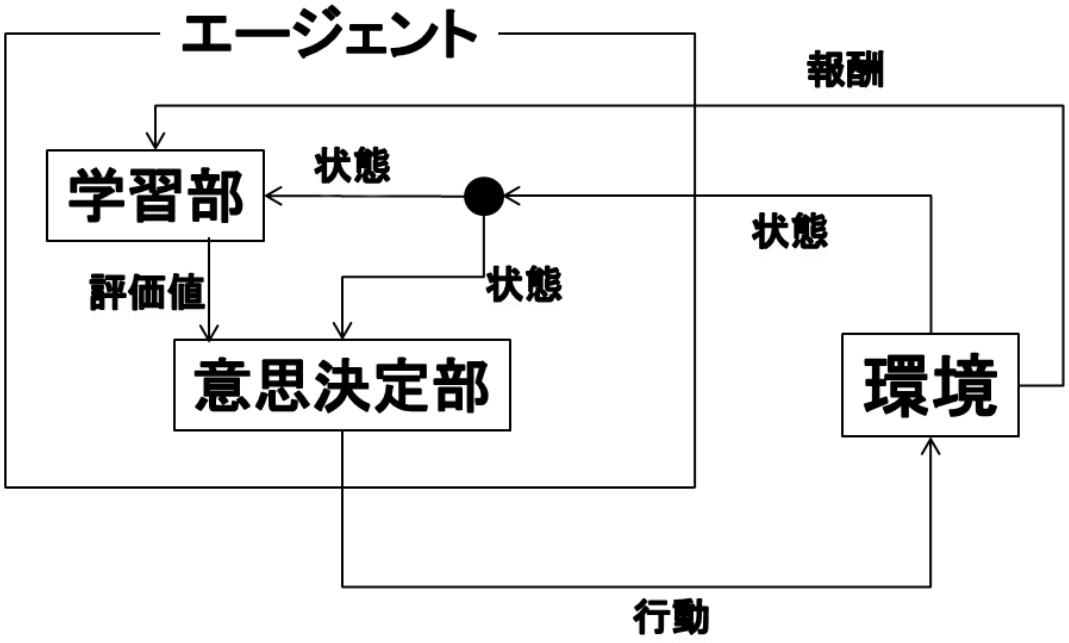


図 3.強化学習の概要

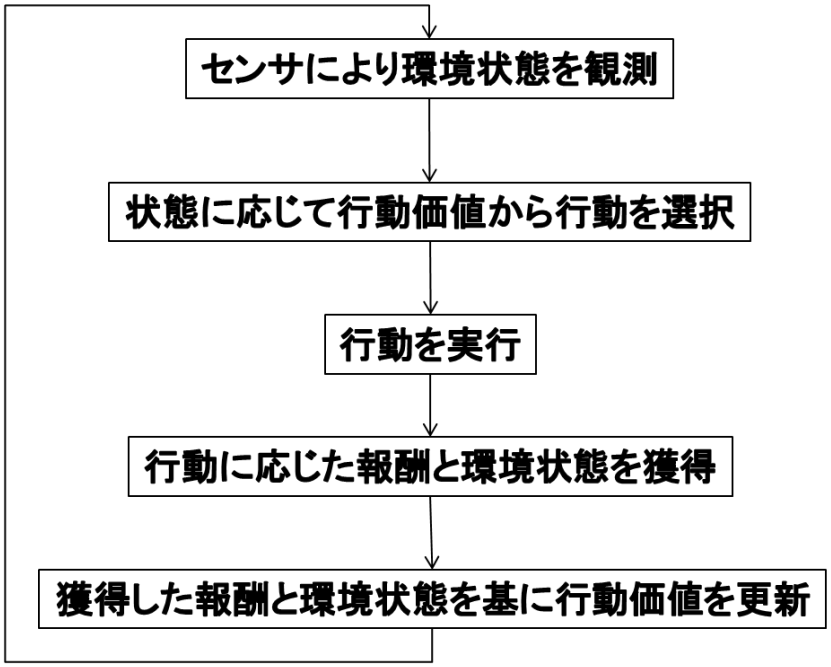


図 4.強化学習での学習の流れ

強化学習は, 教師あり学習と異なり学習のための適切な入力データと出力データのペアが与えられることがない．報酬を基に学習するため，未知の環境でも学習することができる．また，未知の学習領域を開拓していく行動と，既知の学習領域を利用していく行動とをバランス良く選択することができるという特徴も持っている．これは学習を効率よく進めるために行動選択手法を適用することが多いからである．行動選択手法は様々存在する．詳しくは 2.4 節で説明する．強化学習は未知の環境下でのロボットの行動獲得に用いられる[10]．強化学習は環境同定型と経験強化型の 2 つに分類される．

2.2　環境同定型

環境同定型の強化学習では観測した環境に対して最大の報酬を得られる行動を学習する. 環境同定型の強化学習では将来得られる報酬を最大化することを目的としている. つまり, 環境同定型では存在するすべての環境を経験することで最適な方策を取ることが可能となる．　そのため，環境状態を経験していくことで学習を進めていく．環境同定型強化学習はルールを適用したとき, 新環境の行動評価値からそのルールの重みを更新する．本研究では環境同定型の手法である Q 学習を用いている．

Q 学習では行動に対する評価値として Q 値という行動評価値を持つ．Q 値はある環境状態における行動の価値である．環境状態，行動，Q 値の値を軸として構成される空間を Q 空間という．Q 値が高いほどその環境状態で評価が高いつまり報酬を得られるもしくは得るのに近づく行動であるといえる. Q 値はエージェントが行動する度に更新

する．

Q 学習では Q 値を基に行動を選択する．Q 値が正しい評価値となっているなら，行動を選択するとき認識している状態の中で最も Q 値の高い行動をとればよい．しかし，学習を始めた段階では正しい Q 値の値は判明していない．Q 学習は試行錯誤により Q 値の更新を繰り返し， Q 値を最大化するように学習する．

Q 学習の Q 値の更新式について説明する．エージェントが時刻 t において環境状態 s_t で行動 a_t を選択したとする．エージェントは行動実行後，時刻 t＋1 になり環境状態 s_{t+1} に移行する．この時，報酬 r_{t+1} を環境から受け取ったとする．この場合の環境状態 s_t における行動 a_t の評価値 $Q(s_t, a_t)$ は式(1)を用いて更新される．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right] \qquad (1)$$

式(1)における α は学習率，γ は割引率という係数で，それぞれ $0 < \alpha \leq 1$，$0 < \gamma \leq 1$ の範囲で任意に設計者が決定する．$\max_a Q(s_{t+1}, a)$ は環境状態 s_{t+1} における最大の Q 値である．

Q 学習は有限マルコフ決定過程において全ての状態が十分にサンプリングできるようなエピソードを無限回試行した場合，最適な評価値に収束することが理論的に証明されている．実際の問題に対してこの条件を満たすことは困難ではあるが，この証明は Q 学習の有効性を示す要素の 1 つとして挙げられる．

2.3 経験強化型

経験強化型の強化学習では，その時点で経験したルールから行動の価値を推定する．そのため，経験していないルールの価値は考慮しない．できる限り多くの状態を経験しなければならない環境同定型に対して，経験強化型は学習の立ち上がりが早い．欠点としては，環境同定型に比べ収束値が最適解とならないことが多い点が挙げられる．経験強化型の強化学習の代表的な手法としては Profit-sharing[11]がある．

2.4 行動選択手法

強化学習では行動を選択する際に，未知の学習領域を開拓していく行動と，既知の学習領域を利用していく行動とをバランス良く選択する必要がある．そのために行動選択手法を適用することが多い．行動選択手法は様々な方法が存在する．代表的な手法としては，定められた確率 ε でランダムに行動する ε -greedy 法や Boltzmann 分布を用いて行動を選択する softmax 法がある．

2.4.1 ε -greedy 法

ε -greedy 法はある一定の確率 ε でランダム行動を選択し，それ以外の確率で行動評価値の最も高い行動を選択する手法である． ε の確率でランダムな行動を取ることで，現在最も高い評価値となっている行動よりも更に良い行動があるかどうかを調べることができる．

本研究では ε -greedy 法を用いている．ε -greedy 法はシンプルかつ一般的に用いられている手法であり，一定確率で必ず探索行動を行うことができるからである．

2.4.2 softmax 法

softmax 法は，行動の評価値に応じて行動を選択する確率を設定し，その確率を基に行動を選択する手法である．この評価値に応じた行動の確率を行動確率と言う．softmax 法では，行動価値の高い行動には最も高い行動確率が与えられる．その他の行動は，行動価値の高い順に行動確率が与えられる．

softmax 法での行動確率$\pi(s,a)$は式(2)で与えられる．$Q(s,a)$は環境状態sにおける行動aの評価値，T は温度と呼ばれる正定数である．P は環境状態sにおいてとれる全ての行動の集合である．

$$\pi(s,a) = \frac{e^{Q(s,a)/T}}{\sum_{p \in P} Q(s,p)/T} \qquad (2)$$

T の値が高い場合には，全ての行動がほぼ同程度の確率で選択され，温度 T が低い場合には行動価値の高低による選択確率の差が大きくなる．

第３章　ロボットにおける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．まず一般的なマルチエージェントシステムの概要について述べる．次にマルチエージェントシステムを用いた強化学習について説明する．最後に本研究で提案する手法のアプローチについて説明する．

3.1　マルチエージェントシステムの概要

自ら考え行動し、人間に変わって複雑な作業などを代行するものをエージェントと呼ぶ．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自律的に行動し，問題を解決するシステムである．マルチエージェントシステムでは各エージェントはそれぞれに問題を持ち，それを各々解決しながら，システム全体の問題を解決する．マルチエージェントシステムの概要を図 5 に示す．マルチエージェントシステムの特徴は，複数のエージェントから構成される分散システムにある．エージェントの分散環境も空間的分散，時間的分散，意味的分散，機能的分散というように様々な分類が存在する．空間的分散はエージェントが地理的に異なる場所に存在するものを意味する．時間的分散は時間的にずれて発生するものを指す．意味的分散は異なった言語体系やオントロジーで利用するものを意味する．機能的分散は異なった知覚，行動，認知的な能力で利用するものを意味する．

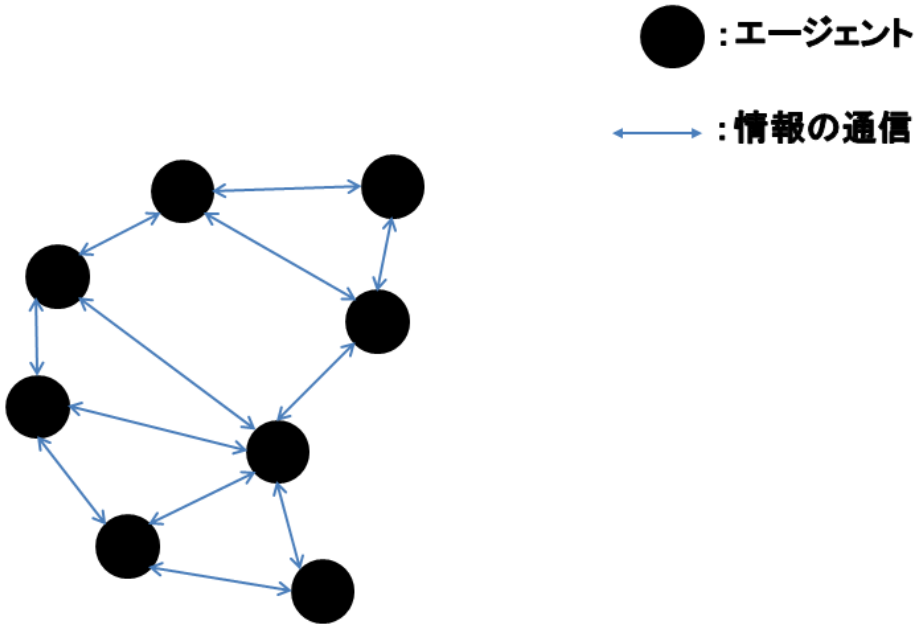


図 5.マルチエージェントシステム概要

マルチエージェントシステムの利点として挙げられるのは分散処理によるシステムの柔軟性がある点である．マルチエージェントシステムにおいて，各エージェントはそれぞれの意思で行動しつつも，全体としての問題を解決する．複雑な問題であればある程，エージェント 1 つで構成されるシステムに比べて，分散処理を行うマルチエージェントシステムの方が柔軟に対応できる可能性が高いのである．マルチエージェントシステムには他にもロバスト性，並列性，モジュール性などの利点がある．

マルチエージェントシステムに関する研究は様々な分野に渡っている．人工知能分野だけでなく，経済学などで広く扱われている．本研究ではマルチエージェントシステムを用いた強化学習に関する研究を扱っている．

3.2 マルチエージェントシステムを用いたロボットの強化学習

マルチエージェントシステムを強化学習に用いることにより，エージェント 1 つで行う強化学習では解決が難しい問題を解く事が出来る．マルチエージェントシステムを用いた強化学習をロボットに適用した研究[12-14]がある．これらの研究は 1 つのロボットに 1 つのエージェントを持たせ，そのロボットが複数存在する環境で強化学習を適用するものである．

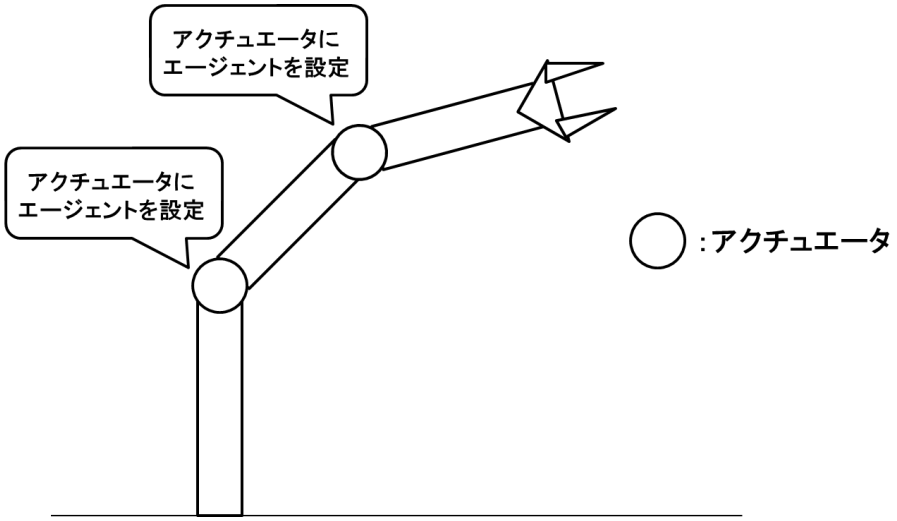
本研究で扱うマルチエージェント環境は 1 つのロボットに対して複数のエージェントを設定するものである．そのため，先にあげた研究の手法をそのまま用いることは不可能である．理由としては各エージェントの与える影響が大きく異なることが挙げられる．

第４章　提案手法

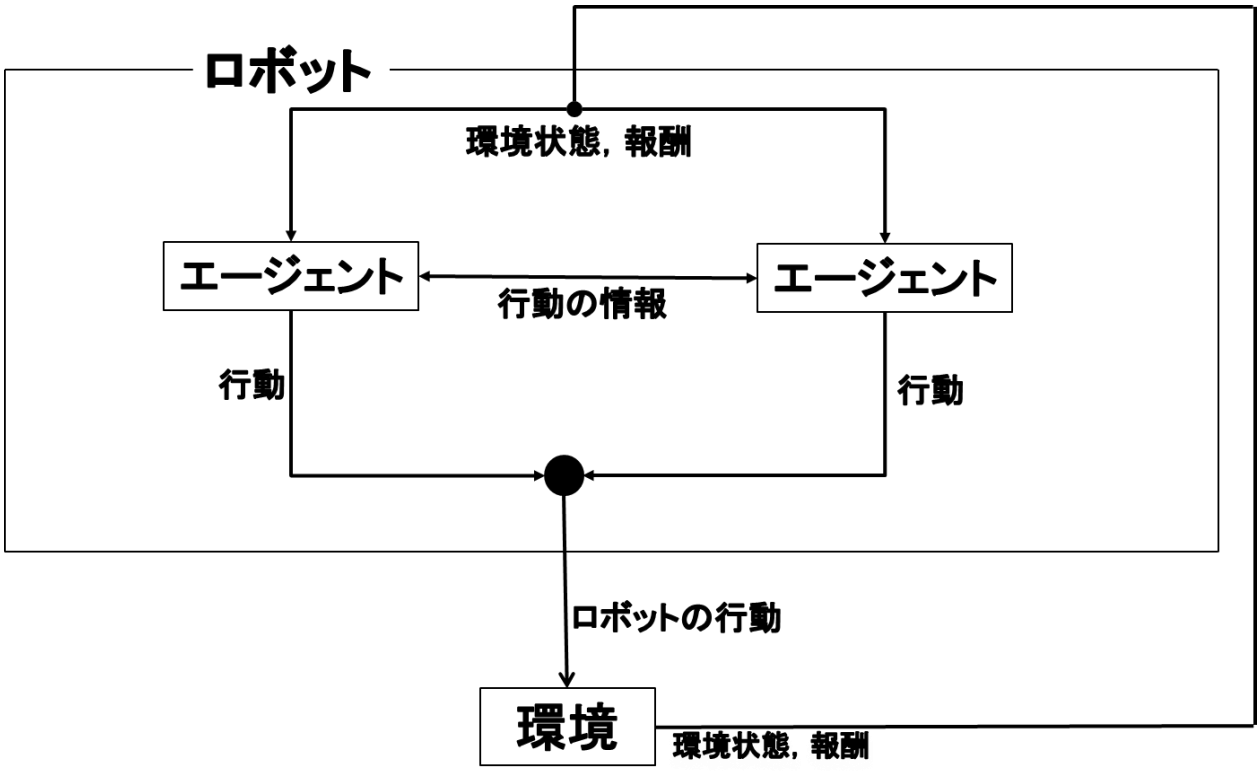
本章では提案手法について説明する．まず，本研究で使用するマルチエージェントシステムについて説明する．その次に提案手法の概要を説明する．そして，提案手法の詳細と流れを説明する．

4.1　本研究で用いるマルチエージェントシステム

本研究で用いるマルチエージェントシステムは, 単体のロボットに対して適用するものである．本研究で用いるマルチエージェントシステムの概要を図 6 に示す．本研究で用いるエージェントは強化学習エージェントである. ロボットに搭載されているアクチュエータ 1 つにつき 1 つのエージェントを設定する. 各エージェントは対応するアクチュエータの行動を学習する. 各エージェントの環境状態はロボットが環境から観測する共通のものと，各エージェントから見た他エージェントの行動の情報で構成される．



(a)エージェントの設定



(b)システム全体の概要

図 6.本研究で用いるマルチエージェントシステム

本研究で用いるマルチエージェントシステムでは，各エージェントの行動選択を同時に行う．各エージェントがばらばらに行動選択をし，行動を出力すると，行動選択を終っていないエージェントの環境状態が実質的に変化し正しい行動選択を行えなくなるからである．全エージェントの行動選択が終了した後，各エージェントの選択した行動をロボットの行動として出力する．

4.2 提案手法の概要

シングルロボットの意思決定手法として複数のエージェントを設定し，各エージェントが協調して行動を選択する手法を提案する．本手法では複数のアクチュエータを搭載したロボットを対象とし，ロボットに搭載されているアクチュエータに一対一でエージェントを設定する．本研究ではアクチュエータが n 個搭載されていると考えて，n 個のエージェントを設定する．本手法の概要を図 7 に示す．

本手法では時刻tにロボットの行動選択を行うものとする．ロボットの行動は各エージェントの選択した行動により決定する．本手法において，各エージェントはそれぞれ対応するアクチュエータの動作を学習する．各エージェントは行動評価にロボットが観

測する環境状態に加えて，他エージェントのとった行動の情報を用いる．これにより，他エージェントの行動を考慮した行動選択を可能とする．

本手法では時刻tにおけるロボットの行動選択に対し，各エージェントは複数回行動選択をして最終的な行動を決定する．まず，各エージェントは環境状態を基に行動を選択する．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全エージェントで共有する．各エージェントは共有した他エージェントが選択した行動の情報を環境情報に加える．そして，他エージェントの行動の情報を考慮に入れて再び行動選択をする．本手法では行動の共有と選択を既定のステップだけ繰り返す．既定のステップまで繰り返した時に各エージェントが選択している行動を出力する．

行動の情報共有と行動選択を繰り返していくことにより，各エージェントは他エージェントの行動の情報を蓄積していく．蓄積した行動の情報を基に行動を選択することで他エージェントの選択する行動に合わせた行動を選択することができる．

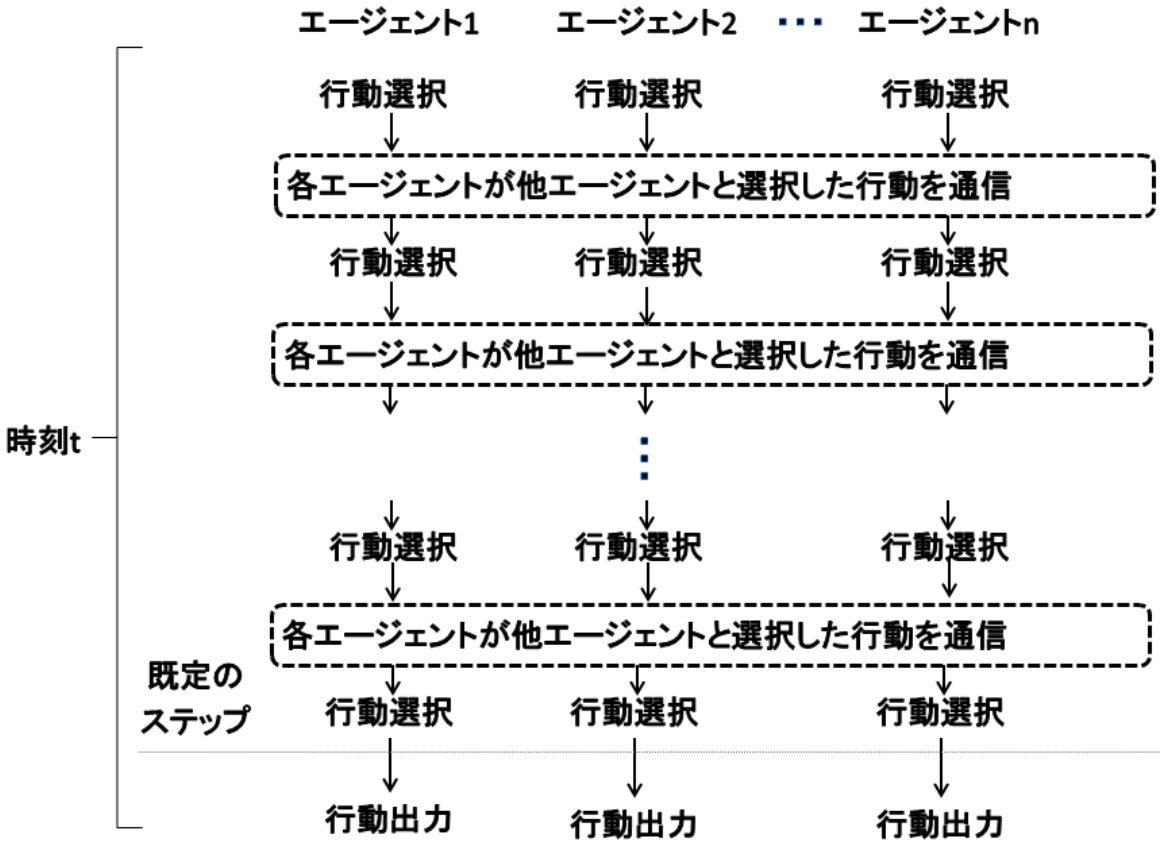


図 7.提案手法の概要

4.3 提案手法の詳細

本手法は単体ロボットの行動獲得に用いる手法である．本手法では各エージェントに 1 から n までの番号を振り分ける．i番目のエージェントをエージェントiと表わす．環境状態をsと表わし，時刻tにおける環境状態をs_tと表わす．エージェントiの行動を行動a_iと表わし，時刻tにおけるエージェントiの行動を$a_{i,t}$と表わす．また，本手法では行動評価値の環境情報に他エージェントの行動の情報を加えている．そのため，各エージェントはロボットが観測する環境状態，他エージェントの行動，自エージェントのとれる行動全ての組に対して行動評価値を持つ．エージェントiにおける他エージェントの行動をA_iと表わす時刻tにおける他エージェントの行動を$A_{i,t}$と表わす．そして，エージェントiのもつ行動評価値を$\mathrm{Q}_i(\mathrm{s}, A_i, a_i)$と表す．特に時刻$t$におけるエージェント$i$の行動評価値は$\mathrm{Q}_i(s_t, A_{i,t}, a_{i,t})$と表される．

本手法での行動選択について説明する．時刻tにおいて，ロボットは環境状態s_tを観測する．ロボットは観測した環境状態s_tを各エージェントに送る．各エージェントは受け取った環境状態s_tを基に行動を選択する．

本手法では時刻tにおけるロボットの行動選択に対し，既定の回数 N 回各エージェントは仮想的に行動選択を行う．ここで，エージェントの仮想的な行動選択をステップと定義する．N ステップが終了するまで各エージェントは選択した行動をロボットの行動として出力しない．ロボットの一回の行動選択の内，最初に行う行動選択を 0 ステップとし，以降 N ステップまで各エージェントは行動選択を行う． 全エージェントの行動選択が終了した後は次のステップに移行する． 各ステップの行動選択の詳細は 4.3 節で説明する．

N ステップまで終了し，全エージェントの行動が決定した後に，ロボットの行動$R_t = \{a_{1,t}, a_{2,t}, \ldots, a_{n,t}\}$を出力する．この流れをロボットの一回の行動選択とする．

4.4 提案手法における行動選択

本手法における各ステップの行動選択では本手法で定義する協調的行動評価値$r(a_i)$を用いて行う．協調的行動価値は強化学習により定義される行動評価値と行動評価値とは別の行動評価指標から算出される．本研究ではその行動評価指標として行動遷移確率$\pi_{a_i}(A_i)$を考える．

行動遷移確率$\pi_{a_i}(A_i)$は自エージェントの行動a_iに対して他エージェントが行動A_iを選択する確率である．行動遷移確率はロボットの行動選択開始時に行動評価値から式(4)で計算される．式(4)は自エージェントの行動と他エージェントの行動の全ての組み合わせに対して行う．現在の環境状態で行動を出力した際に行動評価値が最大のa_iとA_iの組をもっとも高い確率と考える．行動評価値が最大のものが 1 つであれば確率 1 とし，

複数ある場合はそれらを同じ確率とする．行動評価値が最大のもの以外を確率 0 する．Uは行動評価値が最大のものの数である．

$$\pi_{a_i}(A_i) = \begin{cases} \frac{1}{U} & (Q = \max(Q)) \\ 0 & (otherwise) \end{cases} \quad (4)$$

協調的行動評価値$r(a_i)$は行動評価値と行動遷移確率$\pi_{a_i}(A_i)$を用いた式(5)から計算される．式(5)は各エージェントがとることのできる全ての行動に対して適用する．求めた協調的行動評価値$r(a_i)$を用いて各エージェントは行動を選択する．

$$r(a_i) = \sum_{p \in P_i} (\mathrm{Q}_i(\mathrm{s}, p\,, a_i) \times \pi_{a_i}(p)) \quad (5)$$

行動遷移確率は各ステップごとに式(6),(7)を用いて更新される．更新は各ステップの行動選択後に行われる．他エージェントが実際に選択した行動A_iに対しては式(6)を適用し，選択されなかった行動A_iに対しては式(7)を適用する．これにより，実際に他エージェントが選択した行動の確率を高くし，それ以外の行動の確率を低くする．

$$\pi_{a_i}(A_i) \leftarrow \pi_{a_i}(A_i) + \beta\{1 - \pi_{a_i}(A_i)\} \quad (6)$$

$$\pi_{a_i}(A_i) \leftarrow \pi_{a_i}(A_i) + \beta\{0 - \pi_{a_i}(A_i)\} \quad (7)$$

各ステップでの行動選択では 0 ステップと 1 から N ステップで流れが異なる．それぞれのステップでの動作は以下のようになっている．

・0 ステップ

1.行動遷移確率作成

2.行動評価値と行動遷移確率から協調的行動評価値を算出

3.行動を選択

4.選択した行動を他エージェントに送信

・1 から N ステップ

1.行動遷移確率を更新

2.行動評価値と行動遷移確率から協調的行動評価値を算出

3.行動を選択

N ステップの行動選択終了後に選択している行動をロボットの行動として出力する．それまでは各ステップで行動選択を行った後に行動を出力しない．

4.5　提案手法の行動選択の流れ

行動選択の流れを説明する．行動選択の流れを図 8 に示す．本手法では時刻tにおけるロボットの行動選択に対し, 既定のステップ N 回各エージェントは行動選択を行う．ここで，jステップでエージェントiが選択した行動を$a_{i,t}^{j}$と表わし，エージェントiからみた他エージェントの選択した行動を$A_{i,t}^{j}$と表わす．ロボットの一回の行動選択の内, 最初に行う行動選択を 0 ステップとし，以降 N ステップまでエージェントは行動選択を行う．

まずロボットは環境状態s_tを観測する．ロボットは観測した環境状態s_tを各エージェントに送る．次に各エージェントは 0 ステップの行動選択を行う．

0 ステップではまず，各エージェントは行動遷移確率を算出する．次に行動評価値と行動遷移確率から協調的行動評価値を算出する. そして各エージェントは算出した協調的行動評価値を基に行動を選択する．全エージェントが行動選択後，次のステップに移る．

0 ステップ以降の行動選択はすべて共通である．ここではjステップの場合の説明をする．まず各エージェントが$j-1$ステップで選択した行動の情報の送受信を行う．これにより, 各エージェントは他エージェントの行動$A_{i,t}^{j-1}$の情報を得る. その後, $a_{i,t}^{j-1}$と$A_{i,t}^{j-1}$を基に行動遷移確率を更新する．そして，行動評価値と更新した行動遷移確率から協調的行動評価値を算出する. 各エージェントは算出した協調的行動評価値を基に行動を選択する．

N ステップまで終了し，各エージェントの選択している行動$\{a_{1,t}^{N},\ a_{2,t}^{N},\ldots,a_{n,t}^{N}\}$を出力する行動$\{a_{1,t},\ a_{2,t},\ldots,a_{n,t}\}$とし，ロボットの行動$R_t=\{a_{1,t},\ a_{2,t},\ldots,a_{n,t}\}$を出力する．この流れをロボットの一回の行動選択とする．

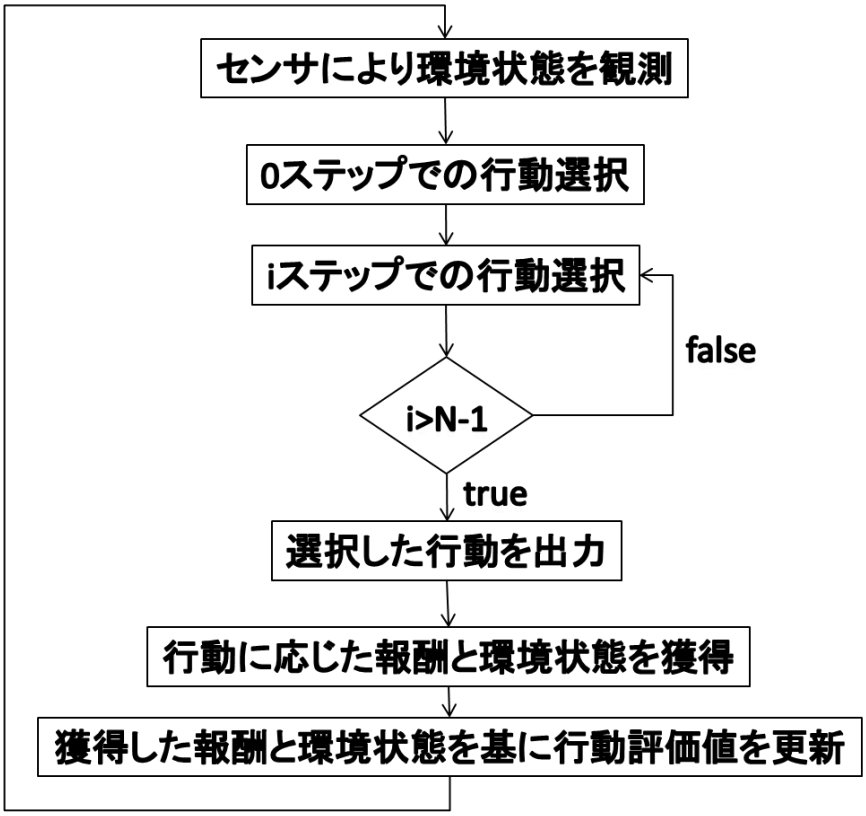


図 8.提案手法の学習の流れ

第５章　実験

本章では本研究で行った実験について述べる．まず実験の目的を述べ，実験の概要を説明する．次に実験の設定を説明する．そして，実験結果を示し，その考察を述べる．

5.1　実験目的

本実験の目的は提案手法の性能を確かめることである．実験はシミュレーションによる検証実験を行う．比較対象はエージェント 1 つで行う従来の強化学習，先行研究のマルチエージェント強化学習，提案手法である．検証内容は提案手法で，タスク達成に必要となる行動を学習し，行動回数が収束した後の試行で，先行研究の手法と比較して最適行動を安定して獲得できるかという点である．

5.2　実験概要

本実験ではロボットアームがリーチング動作によりアームの先端を目標地点に近づくことを目的とするタスクを行う．実験環境を図 9 に示す．本実験は実験環境を二次元平面で表す．本タスクでは多関節のロボットアームで行う．ロボットアームは複数のリンクとアクチュエータで構成されている．アクチュエータはリンクをつなぐ関節に搭載されていて，各アクチュエータは任意の 3 種類の角度に稼動する．ロボットアームは搭載されているアクチュエータを稼働させることでリーチング動作を行う．アームの先端が目標地点に達するとタスクを達成する．タスクを達成すると報酬を得ることができる．

本実験ではロボットアームの関節数を 2 関節，3 関節，4 関節のもので行った．関節数が異なるとロボットとして取れる行動の総数や状態の総数は異なる．そのため関節数を変えて実験を行うことで，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．2 関節の実験を実験 1，3 関節での実験を実験 2，4 関節での実験を実験 3 とする．また 4 関節のものについてはパラメータを変更した実験として実験 4 を行っている．

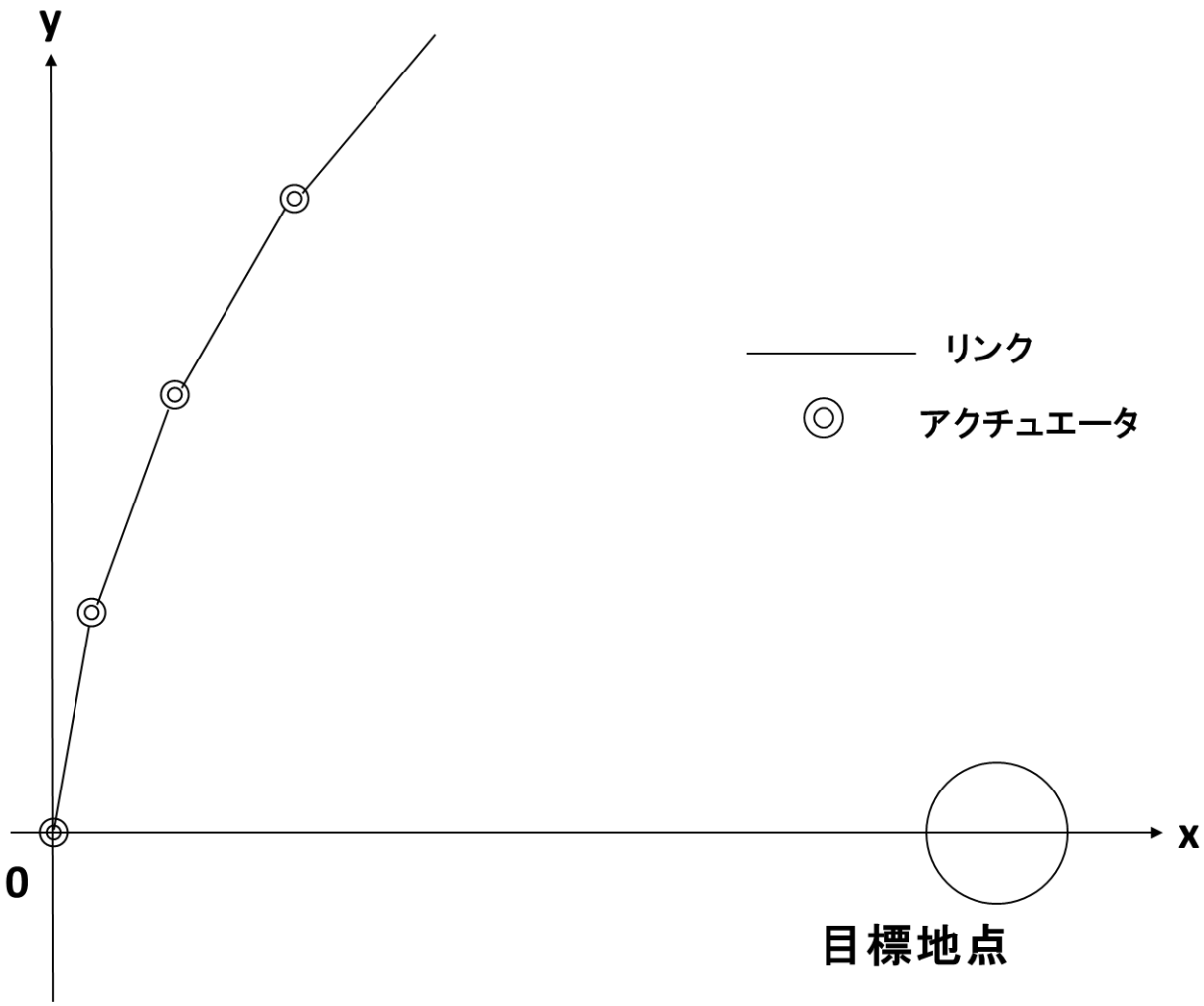


図 9.実験概要

5.3　実験設定

本節では本研究で行った実験の設定を説明する.

5.3.1　ロボットアームの設定

ロボットアームの設定を図 10 に示す. 図 10 では 4 関節の場合を示している. 4 個のアクチュエータの内の 1 個を原点の座標(0,0)で固定する. そのアクチュエータをアクチュエータ 1 とし, その他をアクチュエータ 1 に近い順でアクチュエータ 2, アクチュエータ 3, アクチュエータ 4 とする. 各リンクは原点に最も近いものをリンク 1 とし, 原点に近い順にリンク 2, リンク 3, リンク 4 とする. 各リンクの可動域は全て 0 から 90°である. 各リンクは 1 行動で$-\Delta\theta$°, 0°, $\Delta\theta$° の 3 行動をとることができる. リンク

の設定を図 11 に示す．l_1, l_2, l_3, l_4はリンク 1，リンク 2，リンク 3，リンク 4 のそれぞれの長さである．$\theta_1, \theta_2, \theta_3, \theta_4 (0 \leq \theta_1 \leq 90^\circ, 0 \leq \theta_2 \leq 90^\circ, 0 \leq \theta_3 \leq 90^\circ, 0 \leq \theta_4 \leq 90^\circ)$はリンク 1，リンク 2，リンク 3，リンク 4 がそれぞれの可動する角度である．P_1, P_2, P_3, P_4, P_gはそれぞれアクチュエータ 1，アクチュエータ 2，アクチュエータ 3，アクチュエータ 4，アーム先端の座標である．

P_1 の座標は $(0, 0)$，P_2 は $(x_2(\theta_1), y_2(\theta_1))$，$P_3$ は $(x_3(\theta_1, \theta_2), y_3(\theta_1, \theta_2))$，$P_4$ は $(x_4(\theta_1, \theta_2, \theta_3), y_4(\theta_1, \theta_2, \theta_3))$，$P_g$は$(x_g(\theta_1, \theta_2, \theta_3, \theta_4), y_g(\theta_1, \theta_2, \theta_3, \theta_4))$である．$x_2(\theta_1)$から$y_g(\theta_1, \theta_2, \theta_3, \theta_4)$はそれぞれ式(8)から式(15)で表される．

$$x_2(\theta_1) = l_1 * \cos\theta_1 \tag{8}$$

$$y_2(\theta_1) = l_1 * \sin\theta_1 \tag{9}$$

$$x_3(\theta_1, \theta_2) = x_2(\theta_1) + l_2 * \cos\left(\theta_1 + \theta_2 - \frac{\pi}{2}\right) \tag{10}$$

$$y_3(\theta_1, \theta_2) = y_2(\theta_1) + l_2 * \sin\left(\theta_1 + \theta_2 - \frac{\pi}{2}\right) \tag{11}$$

$$x_4(\theta_1, \theta_2, \theta_3) = x_3(\theta_1, \theta_2) + l_3 * \cos(\theta_1 + \theta_2 + \theta_3 - \pi) \tag{12}$$

$$y_4(\theta_1, \theta_2, \theta_3) = y_3(\theta_1, \theta_2) + l_3 * \sin(\theta_1 + \theta_2 + \theta_3 - \pi) \tag{13}$$

$$x_g(\theta_1, \theta_2, \theta_3, \theta_4) = x_4(\theta_1, \theta_2, \theta_3) + l_4 * \cos\left(\theta_1 + \theta_2 + \theta_3 + \theta_4 - \frac{3\pi}{2}\right) \tag{14}$$

$$y_g(\theta_1, \theta_2, \theta_3, \theta_4) = y_4(\theta_1, \theta_2, \theta_3) + l_4 * sin\left(\theta_1 + \theta_2 + \theta_3 + \theta_4 - \frac{3\pi}{2}\right) \tag{15}$$

3 関節の場合はl_4, θ_4, P_gが存在せず，4 関節におけるP_4がアームの先端の座標となる．2 関節の場合は$l_3, \theta_3, P_3, l_4, \theta_4, P_g$が存在せず，4 関節における$P_3$がアームの先端の座標となる．

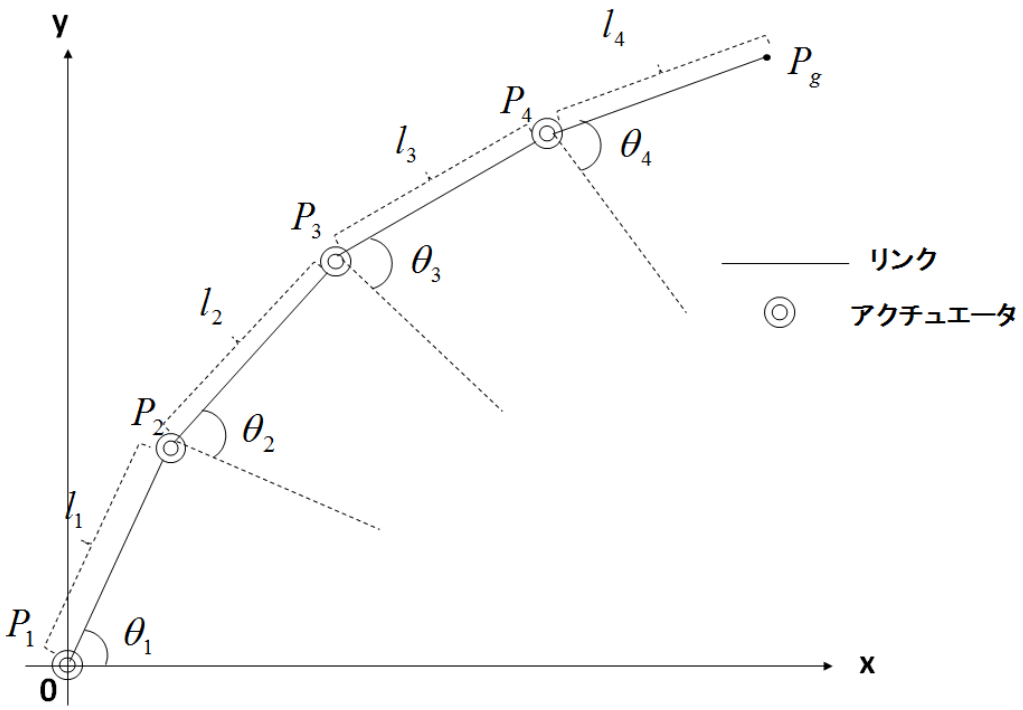


図 10.ロボットアーム設定

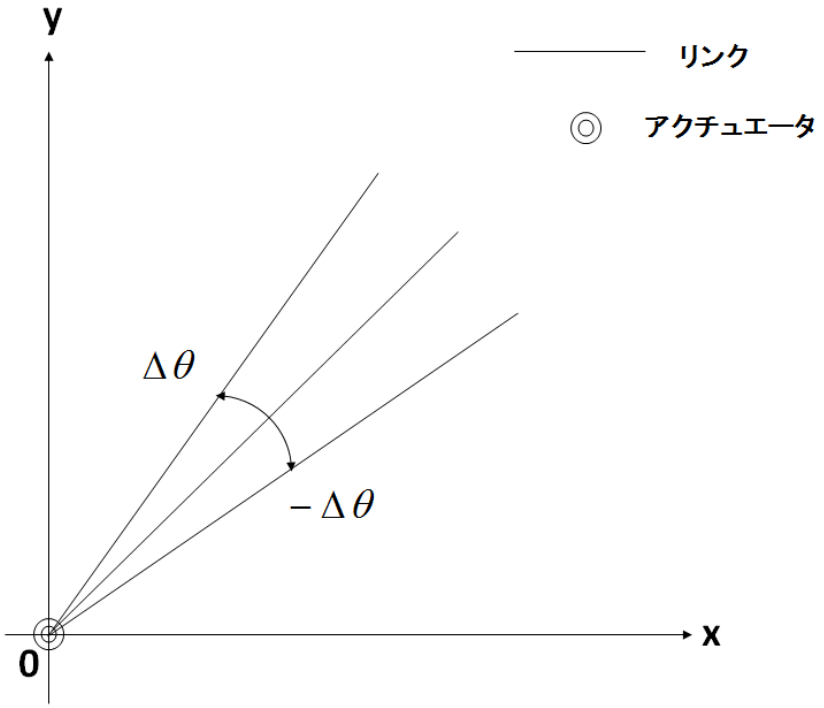


図 11.リンク設定

5.3.2　ロボットアームの状態数と行動数

各リンクはそれぞれ三種類の行動をとることができる. よって, 関節数を n とすると, ロボットが取れる行動は3^nある.

本実験においてロボットが観測する環境状態は各リンクの角度となる. つまり, 総状態数は各リンクの取れる角度の総数の積となる. 今回の設定では全リンク共通で 0°から 90°が可動域となっている. 各リンクの角度の刻みは$\Delta\theta$であるので各リンクの総状態数は$1+(90/\Delta\theta)$で, ロボット全体では$(1+(90/\Delta\theta))^n$状態となる.

表 1 に本実験で想定するロボットアームの 2 関節, 3 関節, 4 関節での行動数と状態数の総数をまとめたものを示す. 本実験では 2 関節, 3 関節のアームは角度の刻みを 5°で, 4 関節は 10°で行っている.

表 1.関節数によるロボットの行動数と状態数の対応

関節数	角度の刻み（°）	総行動数	総状態数
2 関節	5	9	361
3 関節	5	27	6859
4 関節	10	81	10000

5.3.3　目標地点の設定

今回の実験での目標地点の説明をする．目標地点は中心点と中心点の任意の半径の範囲とする．ロボットアームの先端と中心点が一致する条件ではロボットアームの状態数の多さから目的を達成するのが困難だからである．中心点はロボットアームの先端が届く範囲に設定する．目標地点にアームの先端が届けば目的達成とする．目的達成後はロボットアームを初期位置に戻す.

目標地点の場所は各実験により異なる．関節数が異なるとアームの先端が届く範囲が異なるためである.

5.3.4　エージェントの設定

各手法におけるエージェントの設定に関する説明をする. 従来の強化学習, 先行研究, 提案手法の順で説明する.

本実験ではどの手法でもエージェントの学習手法にQ学習を用いている．また行動選択手法には ε -greedy 法を用いている．マルチエージェントシステムにおける強化学習では行動選択手法を適用する範囲が二パターン存在する．1 つ目はロボット単位である. この場合はロボットの一回の行動選択の際にロボット全体で探索行動を行うか最適行動をとるかを決める．つまり, 探索行動をとると決めたら全エージェントが探索行動をとり, 最適行動をとると決めたら全エージェントで最適行動をとることになる．二つ目はエージェント単位である. この場合はロボットの一回の行動選択の際に各エージェン

トがそれぞれで探索行動を行うか最適行動をとるかを決める．つまり，探索行動をとるエージェントと最適行動をとるエージェントが入り乱れることになる．本研究では後者のエージェント単位で行動選択手法を設定する．実験 1～3 では ε =0.05，実験 4 では ε =0.01 に設定している．

従来の強化学習では本実験で用いるロボットアームの行動をエージェント 1 つで学習する．エージェントの学習する行動はロボット全体の行動となる．よってエージェントの行動数は 9 となる．エージェントの環境状態はロボットの総状態数となる．

先行研究のマルチエージェントシステムでは，ロボットアームに搭載されているアクチュエータに対してエージェントを設定する．よって，エージェントはアクチュエータの数だけ設定する．各エージェントの行動数は対応するアクチュエータの行動数となる．本実験ではアクチュエータの行動数は3であるので各エージェントの行動数は3となる．各エージェントの環境状態はロボットの総状態である．

提案手法のマルチエージェントシステムでは先行研究のシステムと同様にロボットアームに搭載されているアクチュエータに対してエージェントを設定する．つまり，各エージェントの行動数は 3 である．提案手法と先行研究での違いは各エージェントの状態数である．提案手法ではエージェントは環境状態として，ロボットの総状態の他に他エージェントの行動の情報をもつ．よって，各エージェントの状態数はロボットの総状態数と他エージェントのとれる行動の総数の積となる．本実験では各エージェントの行動数は等しいため，他エージェントの行動の総数も各エージェントで等しくなる．表 2 に本実験での他エージェントの行動を含めたエージェントの総状態数をまとめる．

表 2.関節数によるエージェントの総状態数

関節数	角度の刻み（°）	他エージェントの行動数	総状態数
2 関節	5	3	1083
3 関節	5	9	61731
4 関節	10	27	270000

各手法によるエージェントに設定する状態数と行動数を表 3 から表 5 にまとめる．

表 3．各手法でのエージェントに設定する行動数と状態数(実験 1)

	従来の強化学習	先行研究	提案手法
行動数	9	3	3
状態数	361	361	1083

表 4．各手法でのエージェントに設定する行動数と状態数(実験 2)

	従来の強化学習	先行研究	提案手法

行動数	27	3	3
状態数	6859	6859	61731

表 5.各手法でのエージェントに設定する行動数と状態数(実験 3)

	従来の強化学習	先行研究	提案手法
行動数	81	3	3
状態数	10000	10000	270000

5.3.5 実験パラメータ

本実験でのパラメータを示す. 関節数や角度の刻みにより異なるパラメータは別々に示す. 表 6 に共通するパラメータを表 7 に実験別のパラメータを示す.

表 6.共通する実験パラメータ

試行回数	10000
報酬	100
学習率 α	0.10
割引率 γ	0.90
ステップ数 N	50
β	0.01

表 7.実験別パラメータ

	実験 1	実験 2	実験 3	実験 4
ε	0.05	0.05	0.05	0.01
θ_1初期値	80	80	80	80
θ_2初期値	80	80	80	80
θ_3初期値		80	80	80
θ_4初期値			80	80
角度の刻み $\Delta\theta$	5	5	10	10
l_1	3	3	3	3
l_2	3	3	3	3
l_3		3	3	3
l_4			3	3
目標地点中心座標	(5.19, 3.0)	(7.79, 4.50)	(12.0, 0.0)	(12.0, 0.0)
目標地点の半径	1	1	1	1

5.4 実験結果

本節では本実験での実験結果を示す.

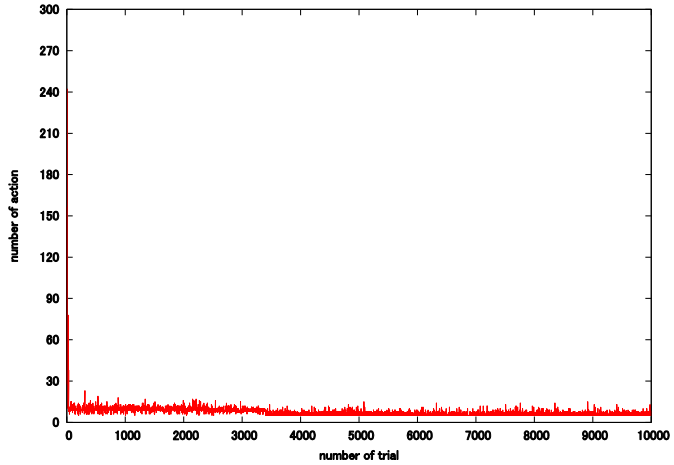
5.4.1 実験 1

本項では実験 1 での実験結果を示す．図 12 に従来の強化学習，先行研究の手法，提案手法の各手法における各試行での行動数を示す．図 13 は各手法での 1 から 1000 試行間での行動数を示す．図 14 には各手法での 9000 から 10000 試行間での行動数を示す．図 15 に各試行までの累計行動数を示す．図 16 に 10 試行ごとの行動数の平均をとったものを示す.

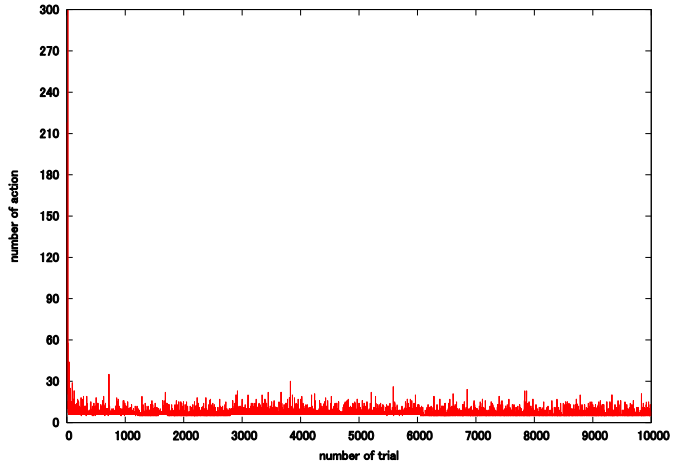
図 12 を見るとどの手法でも早期に学習が収束していることが読み取れる．また，図 13 を見るとどの手法も 100 試行以内に収束していることが読み取れる．従来の強化学習と先行研究の手法では学習初期に行動数が多い試行がある．これは学習初期ではまだエージェントの経験が十分でないため探索行動をとり行動数が多くなっていると考えられる．提案手法では初期の行動数は他の 2 手法と比較すると少ない．これはエージェントが協調して行動を選択するため経験が少なくても目標を達成できているからだと考えられる.

図 14 を見ると，どの手法も行動数 5 が一番少ない行動となっている．このことからどの手法も適切な行動を獲得できていると考えられる．この図から提案手法は先行研究の手法と比較して学習が進んでからの行動数は全体的に少なくなっていることが読み取れる．また，提案手法の学習が進んでからの行動数は全体的に従来の強化学習とほぼ同程度になっている.

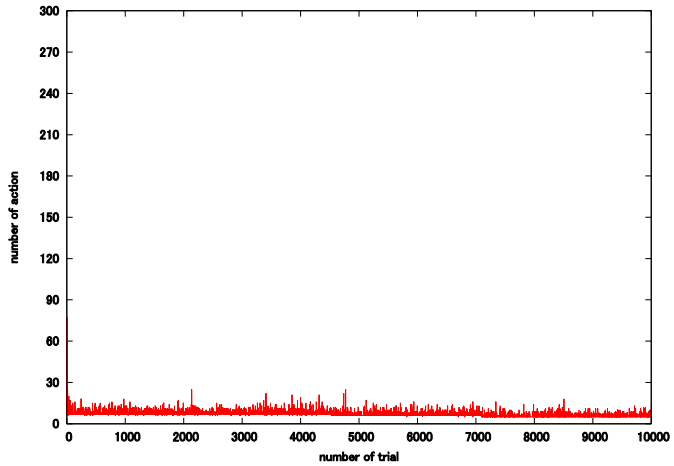
図 15 をみると，累計の行動数は先行研究の手法が一番少なく，従来の強化学習と，提案手法が同程度となっている．図 16 をみると最低行動数 5 に収束するのが従来の強化学習では 350 点目の間，先行研究の手法では約 600 点目，提案手法では約 700 点目の時であることが読み取れる．また，従来の強化学習では 250 点目まで平均値がおよそ 8 から 10，250 から 350 点目での平均値は 7 から 8，350 から 1000 点目での平均値はおよそ 5 から 6 となっている．先行研究の手法では 100 点目までは平均値 6 から 9，100 から 300 点目ではおよそ 5 から 7, 300 から 600 点目では 6 から 8, 600 から 1000 点目では平均値はおよそ 5 から 7 となっている．提案手法では，450 点目までは平均値はおよそ 7 から 8，450 から 700 点目では平均値はおよそ 6 から 7，700 から 1000 点目では 5 から 6 となっている．以上のことから，提案手法は最低行動数を学習するために他の手法より多くの試行がかかっている．図 15 で先行研究の累計行動数が一番少なかったのは他の手法に比べて，行動数の平均値が少なく推移していたからであると考えられる.



（a）シングルエージェント

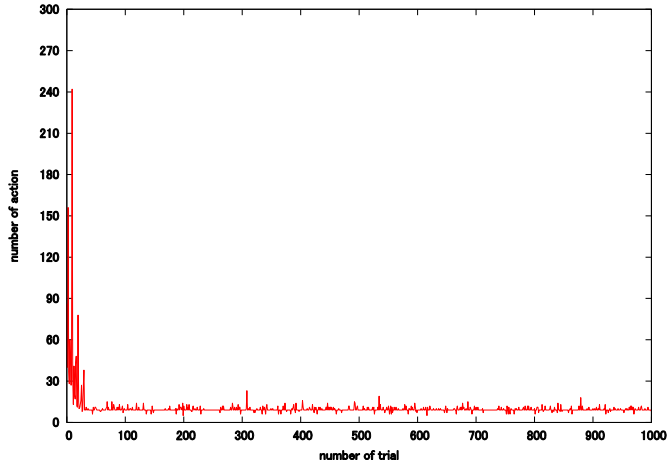


（b）協調なしマルチエージェント

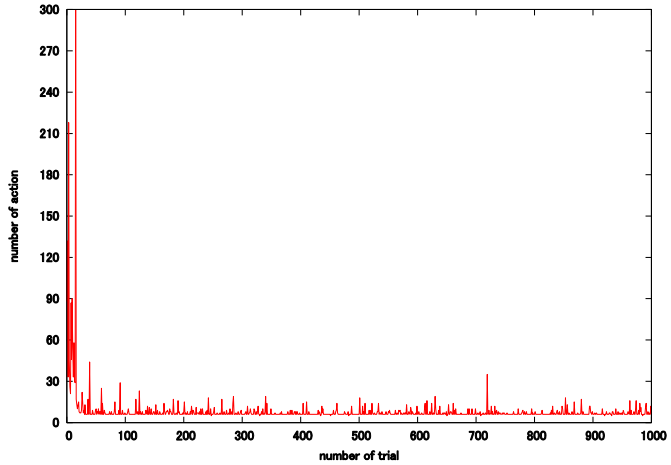


（c）協調ありマルチエージェント[ステップ数 50]

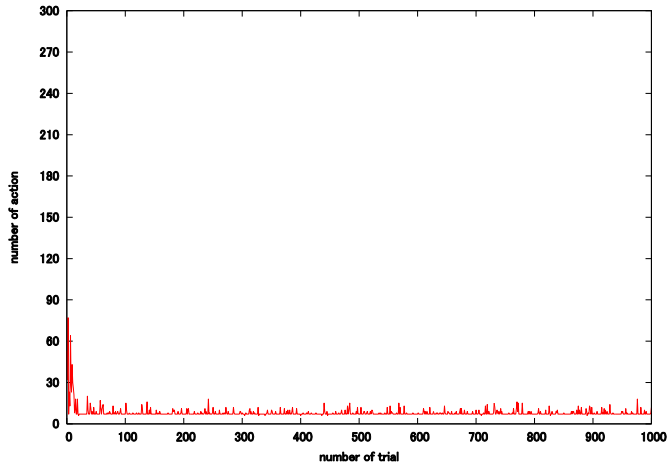
図 12.各試行における行動数



（a）シングルエージェント

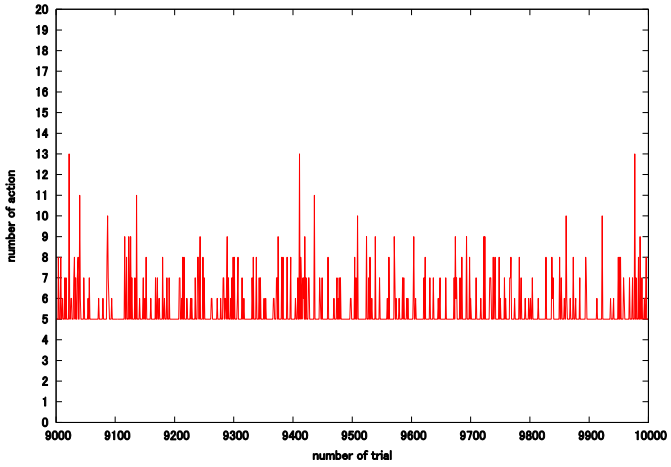


（b）協調なしマルチエージェント

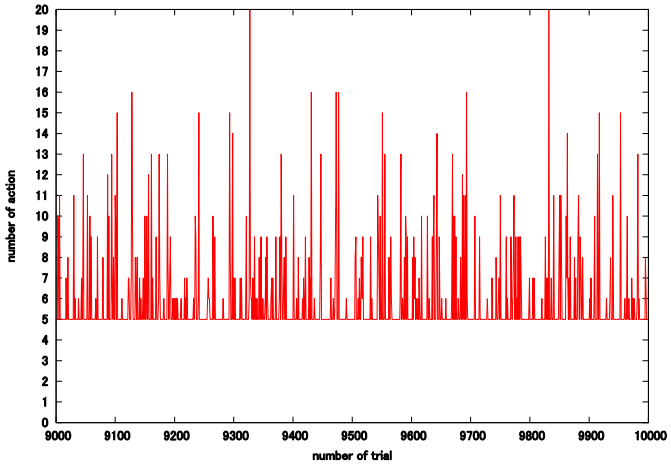


（c）協調ありマルチエージェント[ステップ数 50]

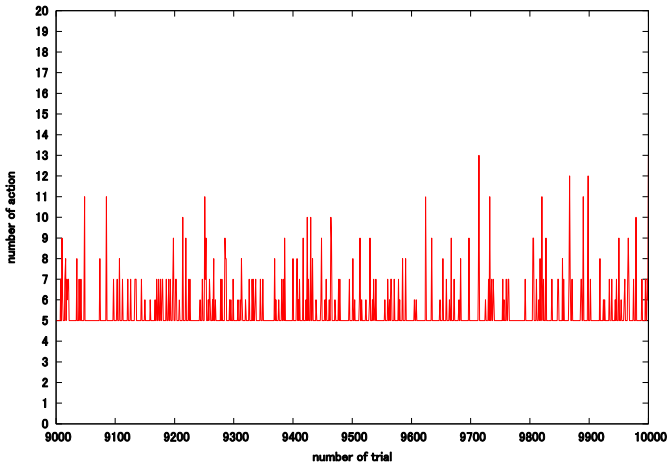
図 13.各試行における行動数(1～1000 試行)



（a）シングルエージェント



（b）協調なしマルチエージェント



（c）協調ありマルチエージェント[ステップ数 50]

図 14.各試行における行動数(9000～10000 試行)

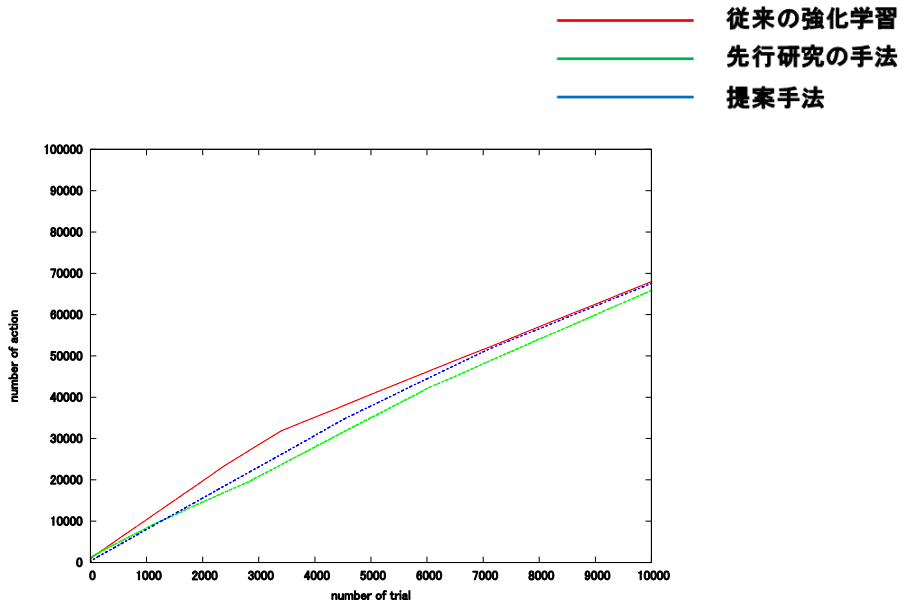


図 15.各試行における累計行動数

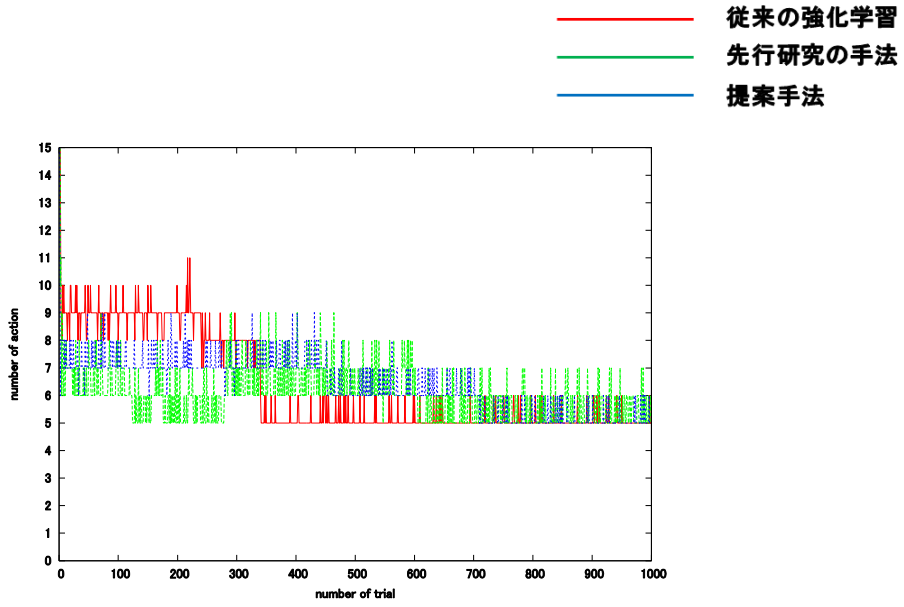


図 16.10 試行ごとの行動数の平均

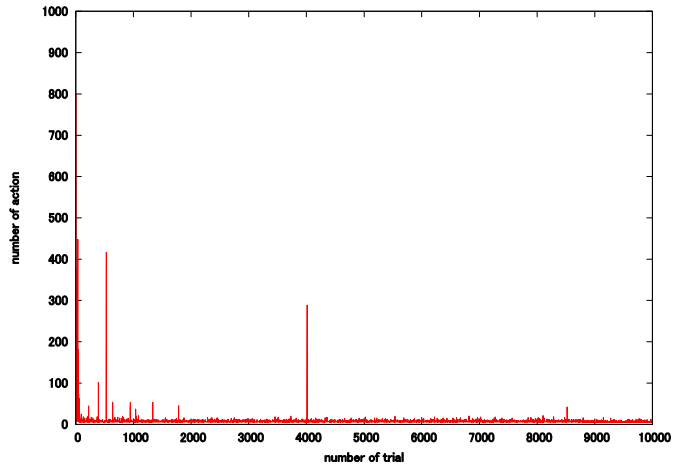
5.4.2 実験 2

本項では実験 2 での実験結果を示す．図 17 に従来の強化学習，先行研究の手法，提案手法の各手法における各試行での行動数を示す．図 18 は各手法での 1 から 1000 試行間での行動数を示す．図 19 には各手法での 9000 から 10000 試行間での行動数を示す．図 20 に各試行までの累計行動数を示す．図 21 に 10 試行ごとの行動数の平均をとったものを示す．

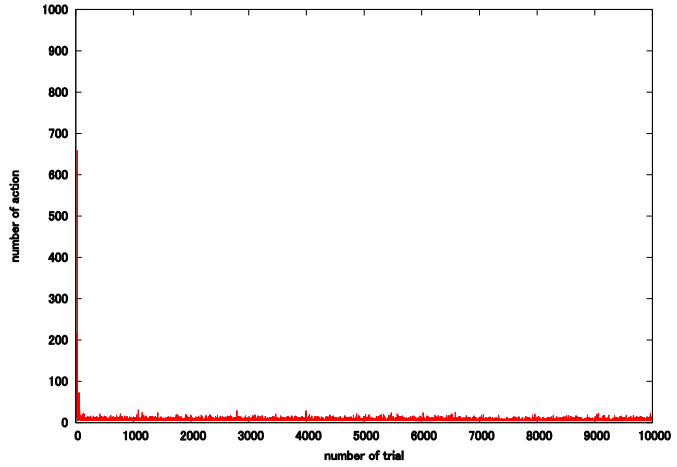
図 17 を見るとどの手法でも早期に学習が収束していることが読み取れる．図 18 を見るとどの手法も 100 試行以内に収束していることが読み取れる．図 18(c)において，提案手法で行動数が異常に多い試行があるが，これはランダム行動により未経験領域に入り込み，行動数が非常に多くなってしまったと考えられる． 図 18(a)の 300 から 400 試行の間と図 18(c)の 600 から 700 試行での行動数が多い試行でも同じことが起こっていると考えられる．

図 19 を見ると，どの手法も行動数 5 が一番少ない行動となっている．このことからどの手法も適切な行動を獲得できていると考えられる．この図から提案手法は先行研究の手法と比較して学習が進んでからの行動数は全体的に少なくなっていることが読み取れる．また，提案手法の学習が進んでからの行動数は全体的に従来の強化学習とほぼ同程度になっている．

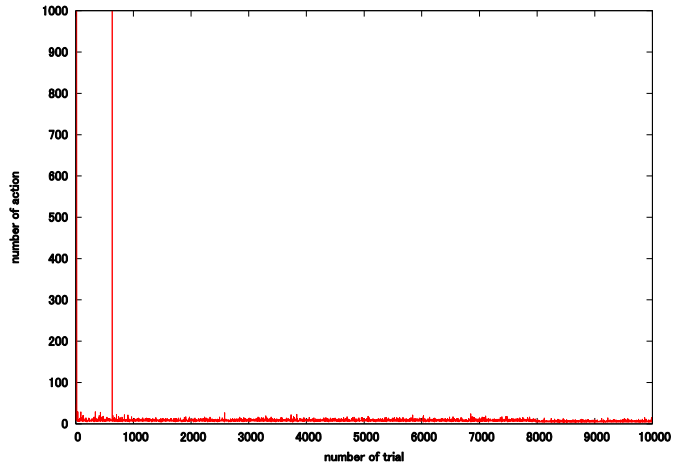
図 20 をみると，累計の行動数は従来の強化学習と先行研究の手法が同程度で，提案手法が一番多くなっている．図 21 をみると最低行動数 5 に収束するのが従来の強化学習では約 900 点目，先行研究の手法では約 600 点目，提案手法では約 800 点目の時であることが読み取れる．また，従来の強化学習では 900 まで平均値がおよそ 6 から 7 であるが非常に平均値が高い点も存在する．900 から 1000 点目での平均値は 5 から 6 となっている．先行研究の手法では 600 点目までは平均値 6 から 8，600 から 1000 点目ではおよそ 5 から 7 となっている．提案手法では，800 点目までは平均値はおよそ 6 から 8，800 から 1000 点目では平均値はおよそ 5 から 6 となっている．以上のことから，提案手法は他の手法より平均値が高い点が多く続いている．これは最低行動数を学習するために他の手法より多くの試行がかかっているといえる．また，今回の結果では図 18(c)から，以上に行動数の多い試行がある．図 20 で提案手法の累計行動数が一番多かったのは他の手法に比べて，行動数 5 をとれる行動を学習しきるのが遅かったことと，行動数が異常に多い試行があったからであると考えられる．



（a）シングルエージェント

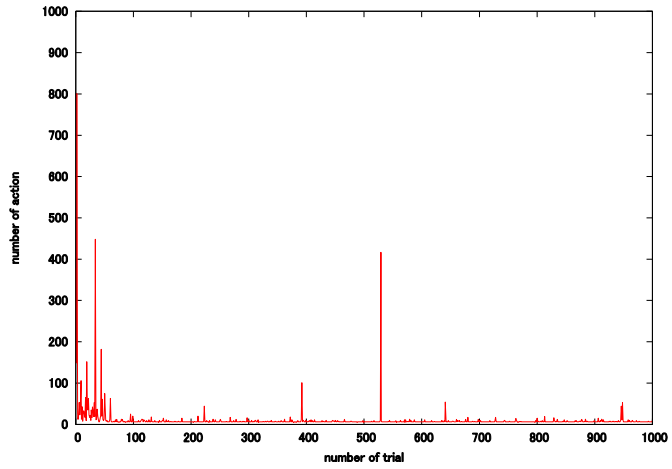


（b）協調なしマルチエージェント

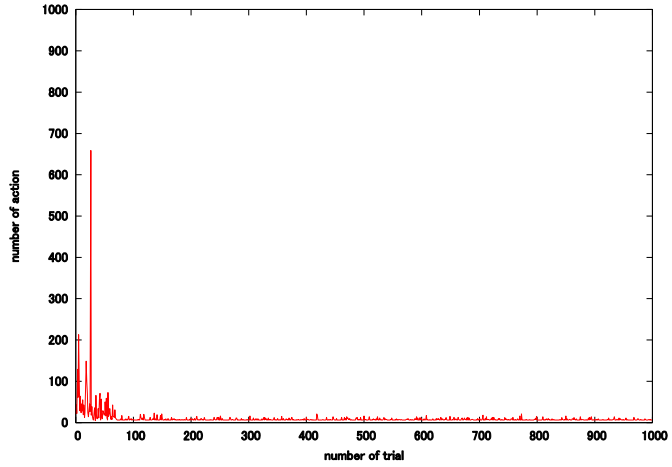


（c）協調ありマルチエージェント[ステップ数 50]

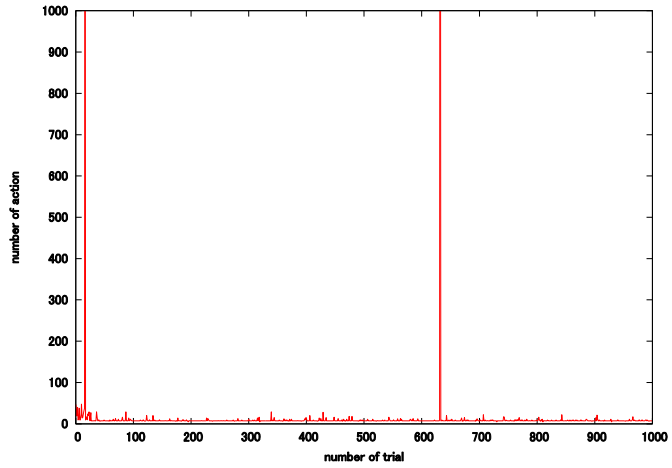
図 17.各試行における行動数



(a) シングルエージェント

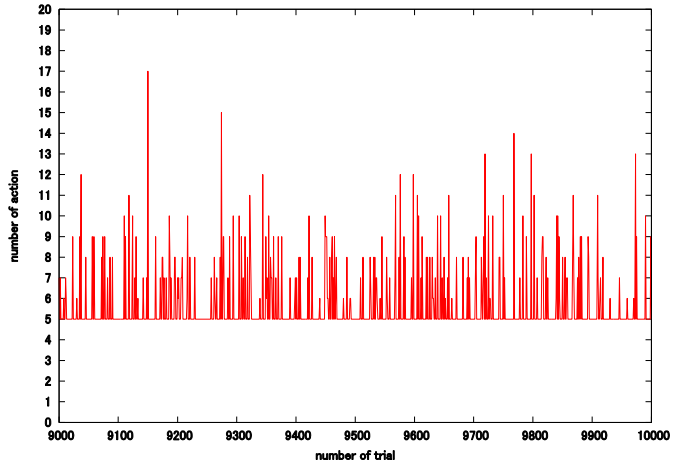


(b) 協調なしマルチエージェント

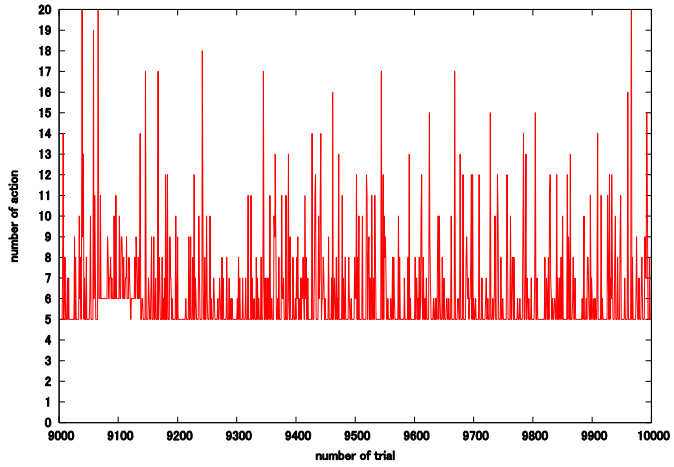


(c) 協調ありマルチエージェント[ステップ数 50]

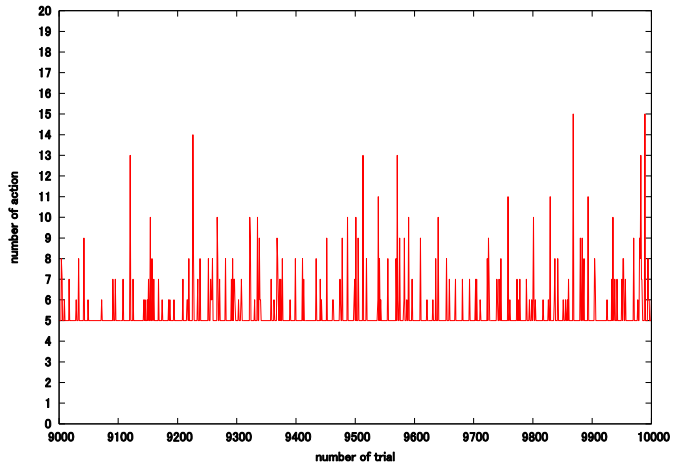
図 18.各試行における行動数(1～1000 試行)



（a）シングルエージェント



（b）協調なしマルチエージェント



（c）協調ありマルチエージェント[ステップ数 50]

図 19.各試行における行動数(9000～10000 試行)

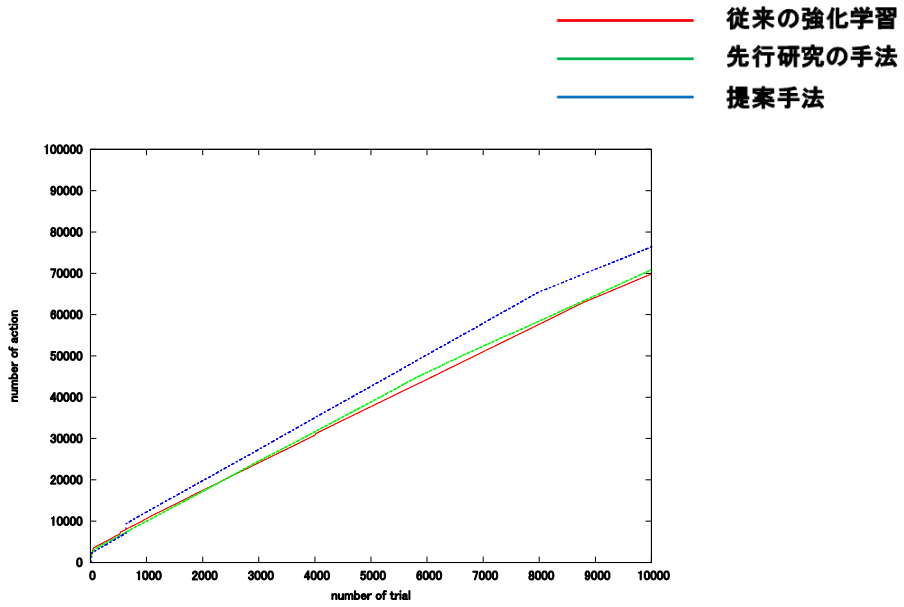


図 20.各試行における累計行動数

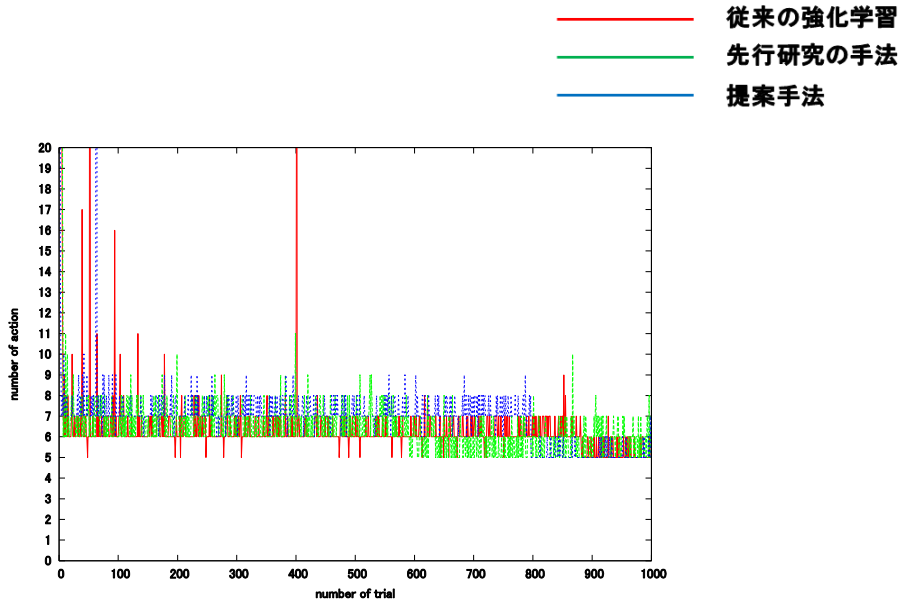


図 21.10 試行ごとの行動数の平均

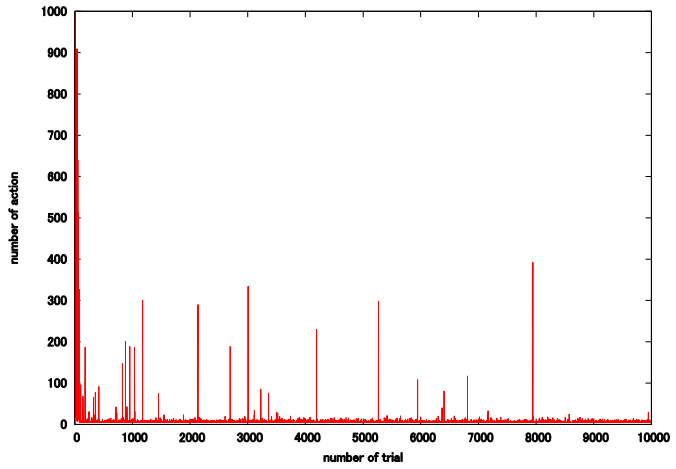
5.4.3　実験 3

本項では実験 3 での実験結果を示す．図 22 に従来の強化学習，先行研究の手法，提案手法の各手法における各試行での行動数を示す．図 23 は各手法での 1 から 1000 試行間での行動数を示す．図 24 には各手法での 9000 から 10000 試行間での行動数を示す．図 25 に各試行までの累計行動数を示す．図 26 に 10 試行ごとの行動数の平均をとったものを示す．

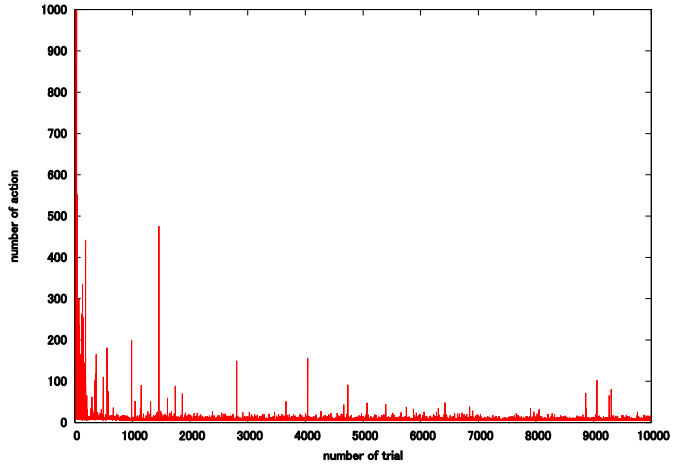
図 22 を見るとどの手法でも学習が進んでからも行動数が 100 を越える試行があることがわかる．これはランダム行動により未経験領域に入り込み，行動数が非常に多くなってしまったと考えられる．4 関節では状態数が多く，ロボットがとれる行動も増えるためこのような結果が表れたと考えられる．特に，先行研究の手法ではエージェントのとりうる行動が計 81 個ある．そのためランダム行動により目標から遠ざかる行動をとり未経験領域に入ると行動数が増えてしまうと考えられる．図 23 を見ると提案手法が一番少ない試行数，行動数で収束に向かっている．図 23 の(b)と(c)を比較すると提案手法の(c)の方が行動数は少なくなっている．

図 24 を見ると，どの手法も行動数 5 が一番少ない行動となっている．このことからどの手法も適切な行動を獲得できていると考えられる．この図から提案手法は先行研究の手法と比較して行動数が全体的に少なくなっていることが読み取れる．これは，提案手法の協調動作により，各エージェントが協調して行動を選択できているからだと考えられる．また，提案手法の学習が進んでからの行動数は全体的に従来の強化学習とほぼ同程度になっている．このことから，提案手法は従来の強化学習と同等の行動を獲得出来ていると考えられる．

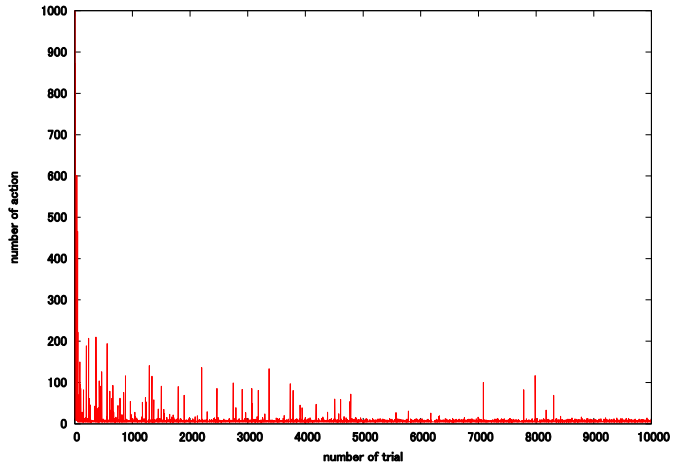
図 25 をみると，累計の行動数は提案手法が一番少なくなっている．その次に少ないのが従来の強化学習で，先行研究の手法が一番多くなっている．この理由は図 25 及び図 23 から，学習初期の試行における行動数が提案手法は少なかったことが影響していると考えられる．提案手法図 26(a)から従来の強化学習では 200 点目つまり 2000 試行程度からは平均行動数がおよそ 5 から 6 となっている．図 26(b)から先行研究の手法ではおおよそ 100 点目つまり 1000 試行程度で落ち着いているがその平均行動数はおよそ 5 から 8 と従来の強化学習に比べ悪くなっている．図 26(c)から提案手法では 500 点目つまり 5000 試行程度からは平均行動数が 5 から 6 となっている．図 26 の結果から，提案手法は従来の強化学習，先行研究の手法と比較すると行動数が収束するために多くの試行が必要になると考えられる．



（a）シングルエージェント

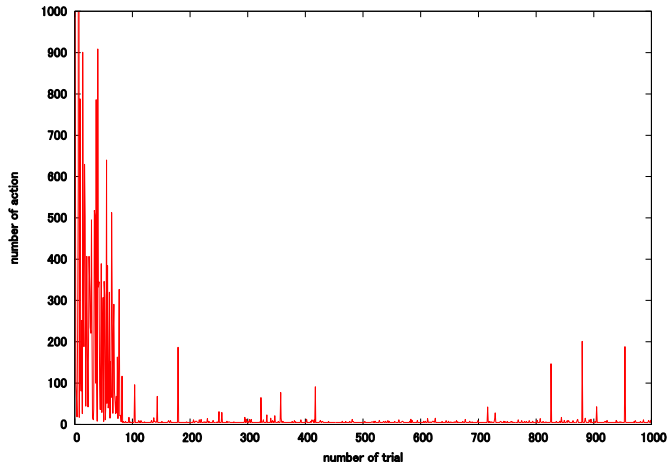


（b）協調なしマルチエージェント

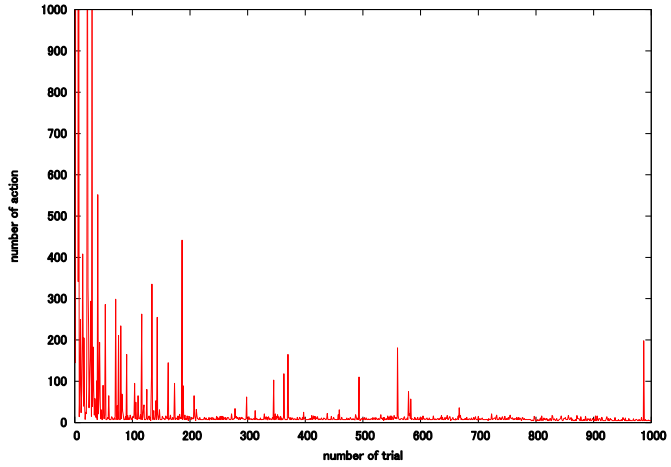


（c）協調ありマルチエージェント[ステップ数 50]

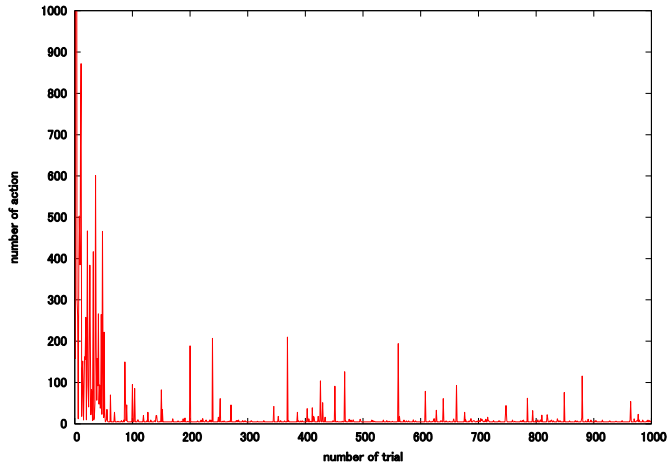
図 22.各試行における行動数



（a）シングルエージェント

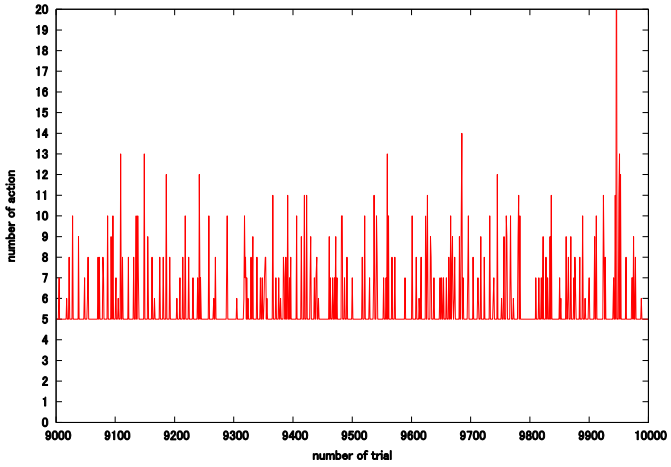


（b）協調なしマルチエージェント

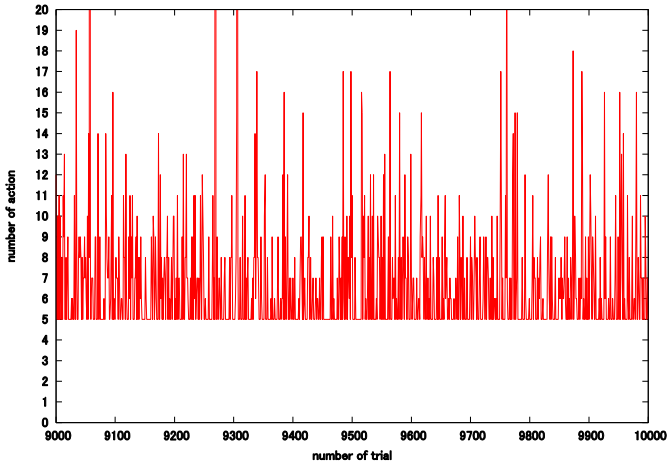


（c）協調ありマルチエージェント[ステップ数 50]

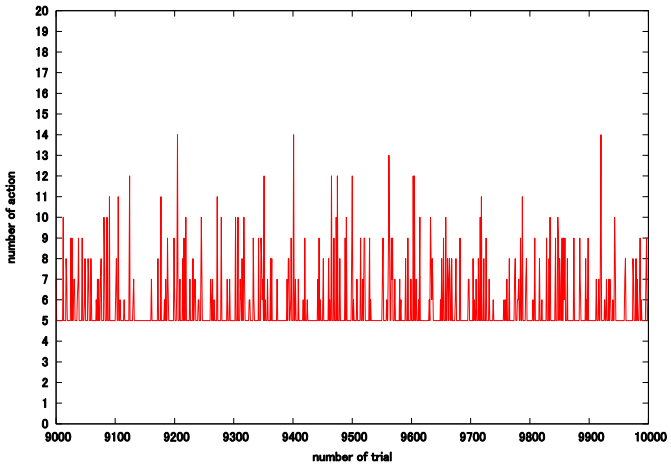
図 23.各試行における行動数(1～1000 試行)



（a）シングルエージェント



（b）協調なしマルチエージェント



（c）協調ありマルチエージェント[ステップ数 50]

図 24.各試行における行動数(9000～10000 試行)

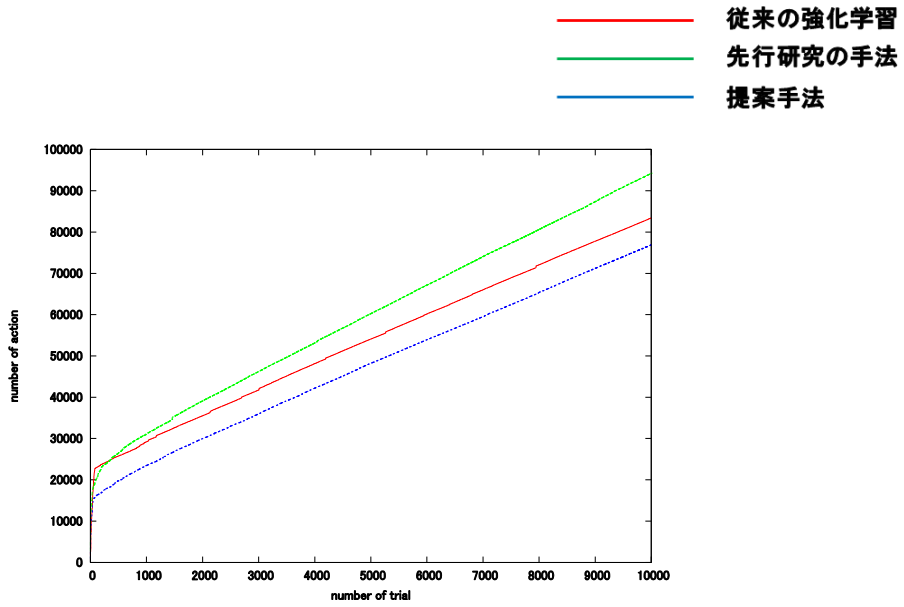
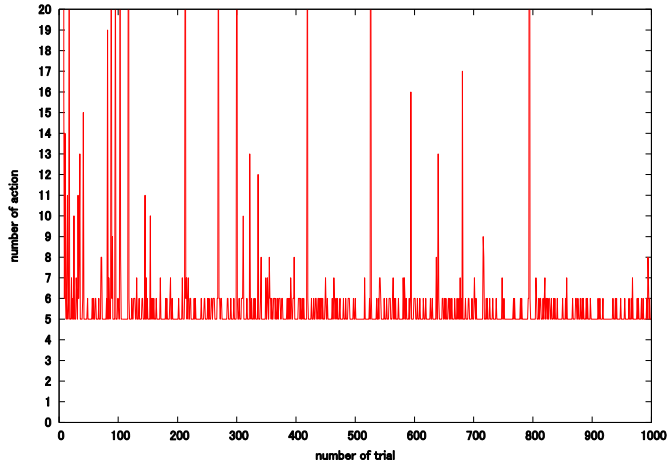
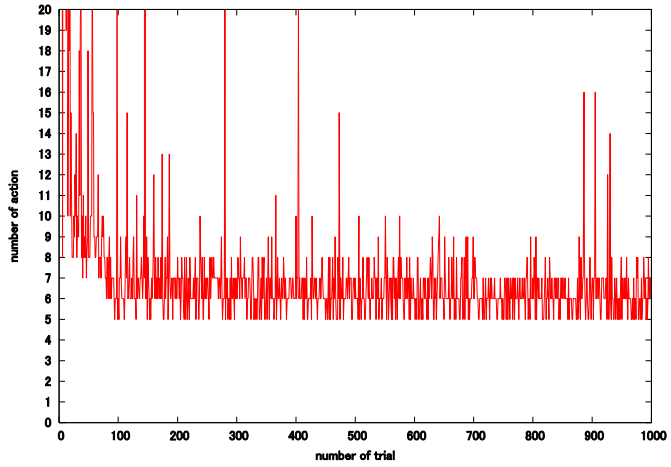


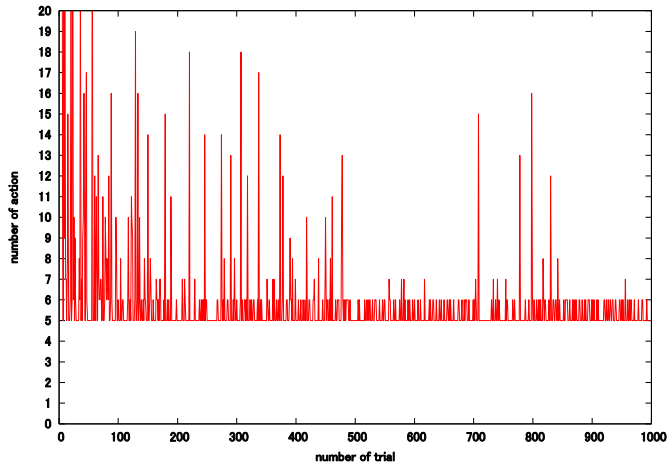
図 25.各試行における累計行動数



（a）シングルエージェント



（b）協調なしマルチエージェント



（c）協調ありマルチエージェント[ステップ数 50]

図 26.10 試行ごとの行動数の平均値

5.4.4 実験 4

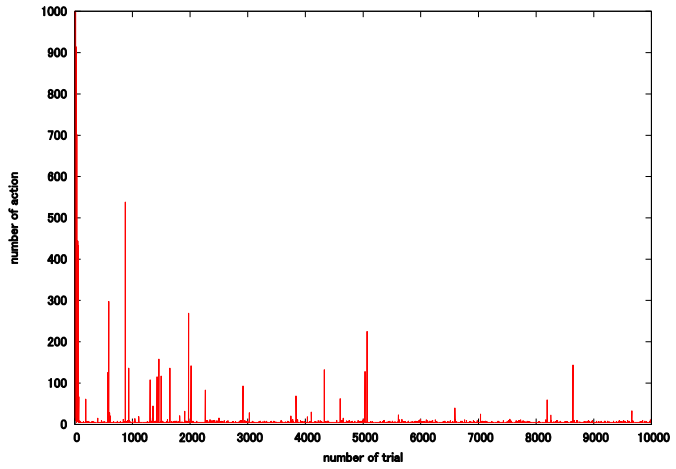
本項では実験 4 での実験結果を示す．図 27 に従来の強化学習，先行研究の手法，提案手法の各手法における各試行での行動数を示す．図 28 は各手法での 1 から 1000 試行間での行動数を示す．図 29 には各手法での 9000 から 10000 試行間での行動数を示す．図 30 に各試行までの累計行動数を示す．図 31 に 10 試行ごとの行動数の平均をとったものを示す．

図 27 を見るとどの手法でも学習が進んでからも行動数が 100 を越える試行があることがわかる．これはランダム行動により未経験領域に入り込み，行動数が非常に多くなってしまったと考えられる．4 関節では状態数が多く，ロボットがとれる行動も増えるためこのような結果が表れたと考えられる．図 28 を見ると先行研究の手法が一番少ない試行で収束に向かっている．

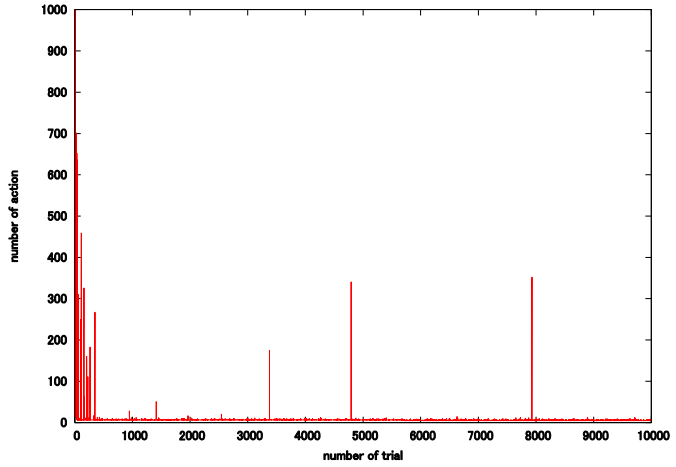
図 29 を見ると，どの手法も行動数 5 が一番少ない行動となっている．このことからどの手法も適切な行動を獲得できていると考えられる．この図から提案手法は先行研究の手法と比較して行動数が全体的に少なくなっていることが読み取れる．これは，提案手法の協調動作により，各エージェントが協調して行動を選択できているからだと考えられる．また，提案手法の学習が進んでからの行動数は全体的に従来の強化学習とほぼ同程度になっている．このことから，提案手法は従来の強化学習と同等の行動を獲得出来ていると考えられる．

図 30 をみると，累計の行動数は提案手法が一番少なくなっている．その次に少ないのが従来の強化学習で，先行研究の手法が一番多くなっている．また，提案手法は 3000 試行程度までは累計行動数は先行研究の手法より多くなっている．しかし，3000 試行を越えてからは先行研究の手法より行動数が少なくなっている．このことから提案手法では学習が進んでからは先行研究の手法より少ない行動数で目的を達成できていることが分かる．

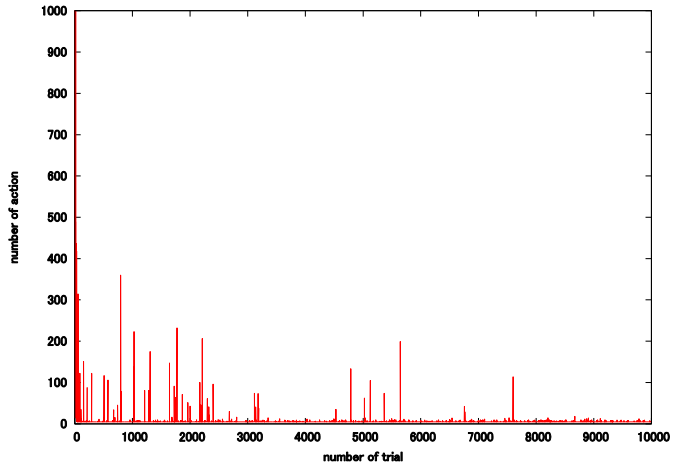
提案手法図 31(c)から提案手法は 600 点目つまり 6000 試行程度からはほぼ行動数が 5 に収束していることが分かる．図 31(a)から従来の強化学習では 500 点目より後つまり 5000 試行以降ではほぼ行動数が 5 に収束していることが分かる．図 31(b)から先行研究の手法では 580 点目つまり 5800 試行程度までは 行動数が 6 で安定し，5800 試行以降に 5 で安定している．このことから行動数の収束に関しては提案手法が一番試行数がかかっていることが分かる．



（a）シングルエージェント

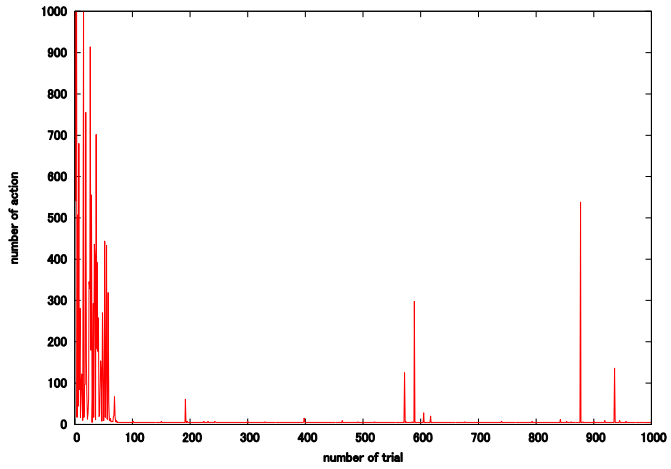


（b）協調なしマルチエージェント

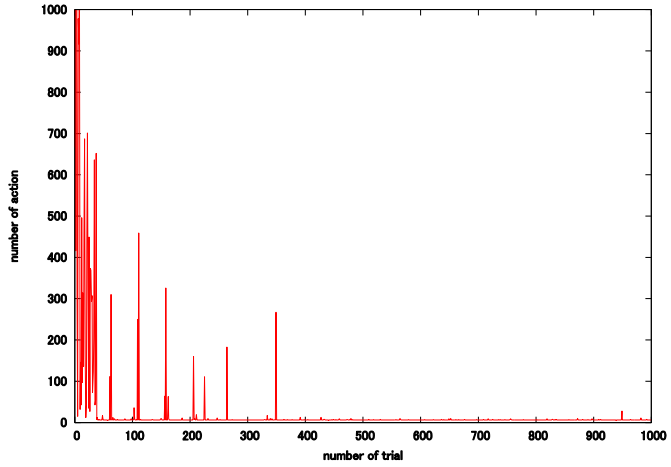


（c）協調ありマルチエージェント[ステップ数 50]

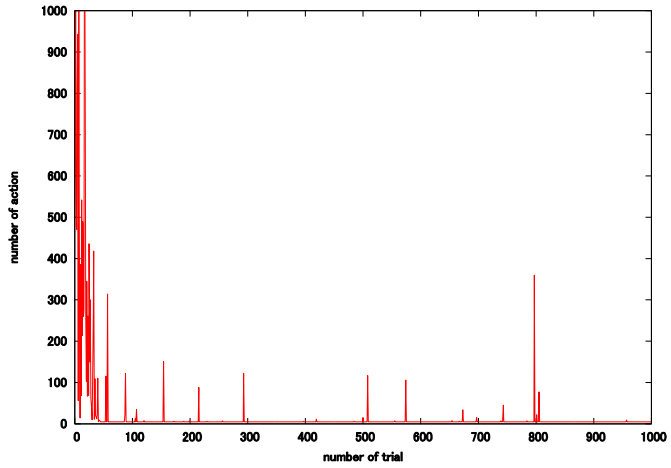
図 27.各試行における行動数



（a）シングルエージェント

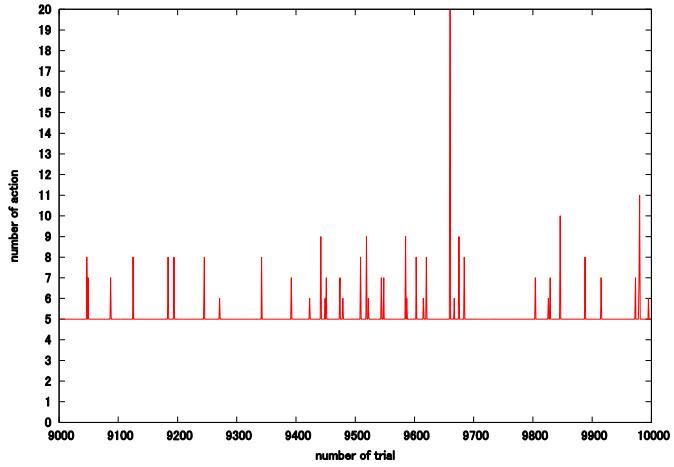


（b）協調なしマルチエージェント

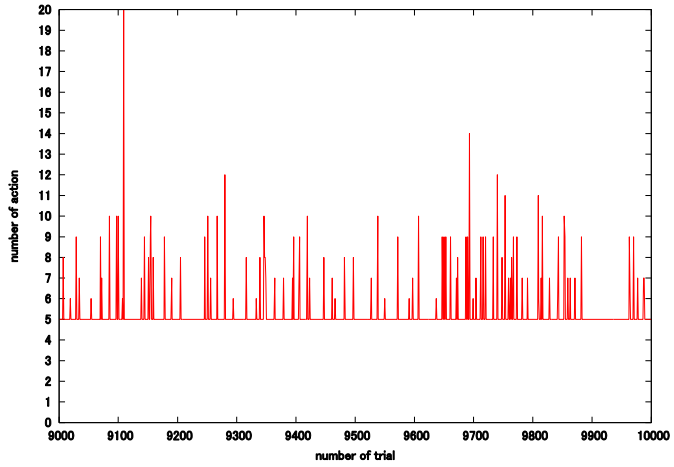


（c）協調ありマルチエージェント[ステップ数 50]

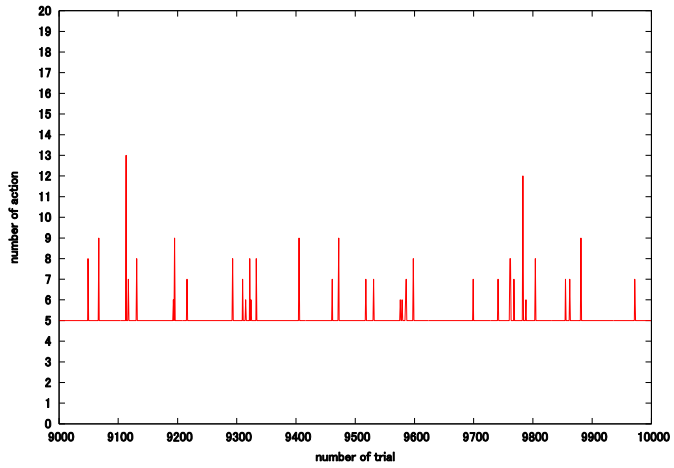
図 28.各試行における行動数(1～1000 試行)



（a）シングルエージェント



（b）協調なしマルチエージェント



（c）協調ありマルチエージェント[ステップ数 50]

図 29.各試行における行動数(9000～10000 試行)

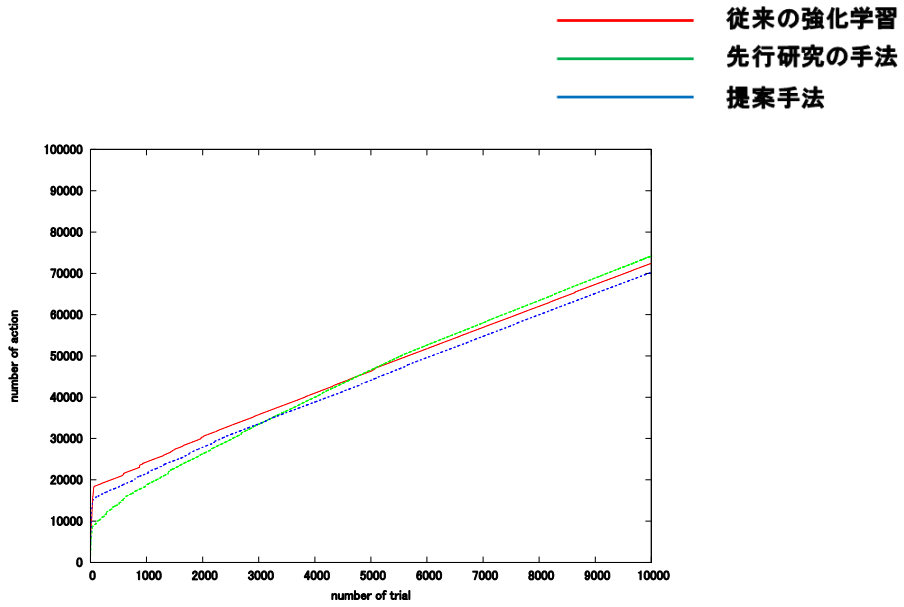
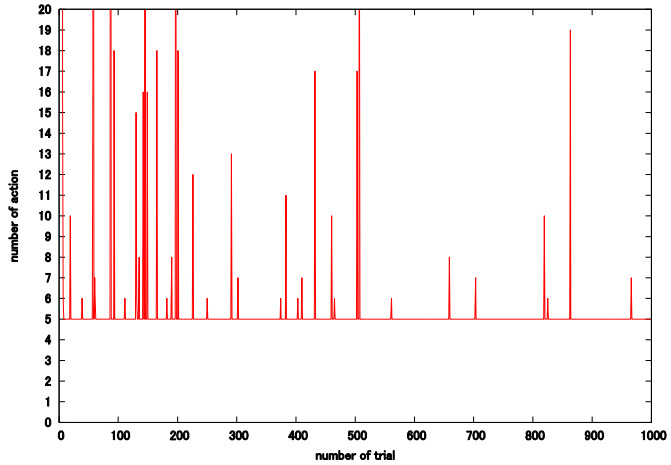
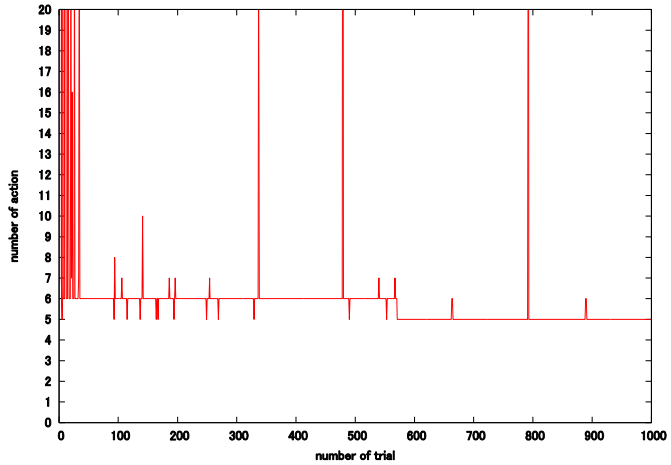


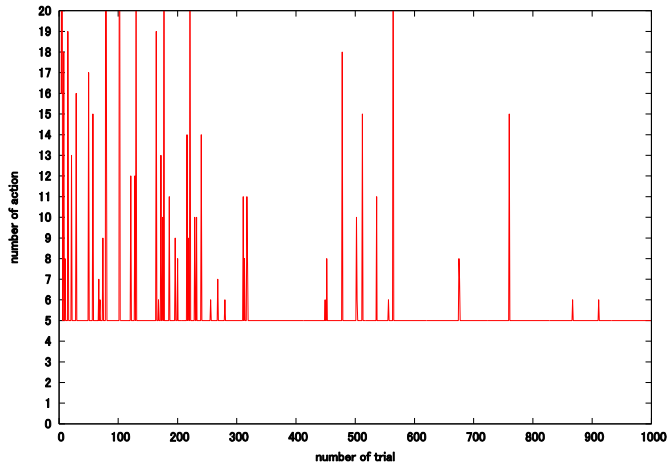
図 30.各試行における累計行動数



（a）シングルエージェント



（b）協調なしマルチエージェント



（c）協調ありマルチエージェント[ステップ数 50]

図 31.10 試行ごとの行動数の平均値

5.5 考察

本実験の実験結果を通しての提案手法の考察を述べる. 各実験における各試行での行動数のグラフから，従来の強化学習，先行研究の手法，提案手法で学習終期での最低行動数が同じであった．このことから，提案手法でタスク達成のための行動を学習できていることが示された.

今回行った実験では，どの実験でも提案手法の収束後の行動数のばらつきが先行研究の手法と比較して少なくなっていた．また，提案手法の学習が進んでからのの行動数は全体的に従来の強化学習と比較して同等の結果となった．これは，提案手法でのエージェント同士の協調が正しくとれており, 各エージェントが他エージェントの行動を考慮した協調行動をとれているからであると考えられる．今回の実験結果から，本研究の目的である先行研究の問題点の学習収束後の行動数のばらつきを抑えることができていると考えられる．よって，提案手法は本研究の目的を達成できている.

一方で, 全試行を通しての累計行動数で 2 関節と 3 関節での実験では提案手法が先行研究の手法より多くなっていた. これは提案手法では目標を最短で達成する行動を学習するのが先行研究の手法より遅いことが原因だった．2 関節の実験では先行研究の手法が目標を最短で達成する行動を学習したのはおよそ 6000 試行で，提案手法では 7000 試行だった．3 関節の実験では先行研究の手法が目標を最短で達成する行動を学習したのはおよそ 6000 試行で，提案手法では 8000 試行だった．4 関節では提案手法の方が累計行動数が少なくなっていたが, これは先行研究の安定後の行動数のばらつきが大きかったのが原因で，学習が安定するのは先行研究の方が早かった．この理由は提案手法のエージェントの協調動作にあると考えられる. 提案手法では他エージェントの行動も考えて各エージェントは行動を選択し学習する．そのため，あるエージェントがランダム行動をとったとしてもその他のエージェントがその行動に合わせて目標達成に近づく行動を選択していく. その結果, ランダム行動による探索をとりにくいと考えられる．このことから，提案手法は局所解に陥りやすいと考えられる．しかし，このランダム行動を抑制する効果は学習が進んでからは安定した行動をとりやすいという利点にもなる.

また今回の実験では行動数が安定してからの行動のばらつきが従来の強化学習と同等の結果となっていた．提案手法では学習の立ち上がりの行動数が少ない．また，提案手法ではエージェントがランダム行動をとったとしても他エージェントがそれを補うことができる．これには探索行動をとりにくいことにもなるが，安定して行動をとることができる. 実際今回の 4 関節の実験では従来の強化学習では学習がある程度安定してからも異常に行動数が多い試行があった. 提案手法でも行動数が多くなる試行もあったが，従来の強化学習よりは抑えられていた．提案手法は従来の強化学習と比較して行動が安定しやすいという点で勝っていると言える.

以上のことから, 提案手法は一体のロボットに複数のエージェントを設定したマルチエージェントシステムにおいて強化学習で各エージェントが適切な行動を学習できることが示された．また，提案手法は他の 2 手法と比較して最適な行動を学習するまで多くの試行を必要とするが, 十分な学習が行われてからは安定した行動をとれることがわかった．このことから提案手法は経験を利用することに優れていると考えられる．

第６章　まとめ

本章では本研究を通しての全体のまとめを述べる．また，本研究の今後の課題についても述べる．

6.1 論文全体のまとめ

本研究では，先行研究の問題点である行動学習を十分に終えてからの行動のばらつきを少なくすることを目標とした．その原因はエージェント間の協調にあると考え，エージェントが協調して行動選択をする手法を考えた．本研究ではエージェントが認識する状態に注目した．エージェントの認識する状態に，他エージェントが選択した行動を加えることで他エージェントの行動も含めて行動を学習できる．各エージェントが他のエージェントが選択した行動を知るためには少なくとも一回は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一回の行動選択に対してエージェントが複数回行動選択を行う手法を考えた．エージェントの複数の行動選択では一回一回行動を出力せずエージェント間で通信を行う．一定回数の行動選択が終わって最終的に各エージェントが選択している行動を出力する．何度も行動選択を行うことでエージェントの行動を一意に定めることができる．

また，本研究ではエージェントの行動評価指標として行動遷移確率を定義した．行動遷移確率はエージェントの行動に対して他エージェントが選択する行動の確率である．行動遷移確率は行動選択する度に更新される．各エージェントが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の組の確率は上げ，選択しなかった行動の組に対しては選択する確率を下げる．本研究において各エージェントは行動選択の際には強化学習で定められる行動評価値と本研究で定義した行動遷移確率を用いて算出される行動評価値を用いる．行動遷移確率を含めた行動評価値を用いることで複数回行動選択により得られる行動の情報を利用して行動選択することができる．

本研究では提案手法により，先行研究の問題点である学習が進んでからの行動数のばらつきを小さくすることを達成できているかと，提案手法の性能を確かめるためにシミュレーション実験を行った．実験内容はではロボットアームによるリーチングタスクを行った．2 関節，3 関節，4 関節の場合に対して実験を行った．　実験結果としては，目的としていた学習安定後の行動数のばらつきを先行研究と比較して少なくすることができた．また，提案手法の特性として予期せぬランダム行動を抑制し行動を安定させることができるが，探索行動をとりにくく局所解に陥る可能性があることがわかった．

6.2 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述べる．

6.2.1 他の実験設定での実験

本研究で行ったシミュレーション実験で想定したロボットアームはどのアクチュエータも可動域が等しく，とれる行動も同じという設定で行った．そのため，全エージェントで状態数ととれる行動の数は同じだった．今後の課題として異なる状態数のエージェントやとれる行動の異なるエージェントでも提案手法の協調動作が問題なく働くかを検証する必要がある．

6.2.2 他の学習手法での検証

本研究で提案した手法では Q 学習を用いて実験を行った．そのため提案手法が適用可能だと証明されているのは Q 学習のみである．行動選択手法に関しても ε -greedy 法しか検証できていない．学習手法については Q 学習以外の手法でも提案手法は有効か検証する必要がある．特に Q 学習は環境同定型の強化学習であるので，経験強化型の強化学習を用いての検証は必要だろう．行動選択手法に関しては他の手法を試すだけでなく，適用する範囲やパラメータについても検討する必要がある．本研究では ε -greedy 法における探索行動を選択するか，最適行動を選択するかの範囲をエージェント単位で行ったが，ロボット単位にするとどうなるかも検証する必要がある．また，エージェント単位で設定するにしても全エージェントで同じ ε の値を用いるのではなく，エージェントごとに ε の値を変えての検証も必要である．

6.2.3 実ロボットへの適用

提案手法は実ロボットへの適用を考えての手法である．本研究で行ったシミュレーションもロボットアームを仮定しての実験であった．しかし，あくまでシミュレーションであり実ロボットに適用しての実験，検証を行っていない．本研究での提案手法を実ロボットに適用した場合，シミュレーション実験では起こり得なかった問題が発生する可能性がある．例えば，シミュレーション実験では各エージェントの同期をとって行動選択を行っているが，実ロボットでは完全に同期がとれない可能性がある．また，エージェント間の通信も実ロボットではタイムラグなどの問題が発生する可能性がある．

実ロボットに適用しての問題は実際に適用してみなければ全てはわからない．実際に実ロボットに適用し問題がでればそこを解決するように手法を改善していく必要がある．

参考文献

[1] 野田五十樹, “ロボットにおける機械学習の課題と動向”, 情報処理 44(11), 2003.

[2] 小池康晴, 鮫島和行, “強化学習の基礎”.

[3] 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 Vo1.13No.1,pp.51~56,1995.

[4] 港隆史,浅田稔, “環境の変化に適応する移動ロボットの行動獲得”, 日本ロボット学会誌 Vol. 18 No. 5, pp.706～712, 2000.

[5] 北村泰彦, “マルチエージェントによる分散協調問題解決”, 計測自動制御学会関西支部 30 周年記念講習会, 1996.

[6] 伊藤孝行,新谷虎松, "マルチエージェントシステムのための実装技術とその応用", 人工知能学会誌 16(4), 469-475, 2001.

[7] 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.

[8] 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.

[9] 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 回記念学術講演会論文, 2012.

[10] 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] 宮崎和光,木村元,小林重信, “Profit Sharing に基づく強化学習の理論と応用”, 人工知能学会誌, 14(5), 800-807, 1999

[12] 内部英治,浅田稔,細田耕, "複数の学習するロボットの存在する環境における協調行動獲得のための状態空間の構成", 日本ロボット学会誌, 20(3), 281-289, 2002

[13] 内部英治,浅田稔,細田耕, "マルチエージェント環境における行動学習のための部分空間同定法による状態空間の構成", 情報処理学会研究報告 ICS, [知能と複雑系] 98(24), 27-32, 1998

[14] 舘山武史,川田誠一,下村芳樹, "探索エージェントを導入した学習経験を共有するマルチエージェント強化学習システムの提案", 日本機械学會論文集 C 編 74(739), 692-701, 2008

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杉本大志さん，高泉昇太郎さん，渋谷和さん，三浦丈典さん，二階堂芳さん，片山和宣さん，小橋遼さんに感謝します．