平成　２３年度

卒 業 研 究 論 文

題　　目　　マルチエージェントによるシングルロボットの行動学習に関する研究

提 出 者　　室蘭工業大学 情報工学科

氏　名　　高泉昇太郎

学籍番号　2023049

提出年月日　　平成　24年　2月　13日

室蘭工業大学
情報工学科

目次

第 1 章　序論

1.1　ロボット研究の概要と従来のロボット制御

ロボットの研究が進むにつれてその制御方法は進歩している．ロボットが社会に登場し始めたのは 1950 年代から 1960 年代にかけての頃である．この時代に登場したロボットは工場などで使用される産業用ロボットである．産業用ロボットとは人間の代わりに作業や保守点検を行うロボットである．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，その動作を何度で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．産業用ロボットが使用される場所は単純な繰り返し作業が主であった．同じ作業を繰り返す点に関しては人間が行うより効率が良い．しかしこの時点のロボットはまだ複雑な作業を行うことはできなかった．

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が可能となった．その大きな理由としてロボットに搭載する感覚機能が開発されたことである．この感覚機能をセンサと呼ぶ．センサとは自然現象や人工物の性質やその空間情報・時間情報を，何らかの科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．センサをロボットに搭載することで，ロボットは自身の現在の状態を数値情報として認識することができる．その数値に応じて自身の行動を選択，変化させることが可能となった．ロボットが自己の行動を制御可能となったことで，ロボットの使用される場面も更に増大した．

現在では，何らかの方法で得た知識を利用して学習を行い，自身の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，事前の人間による設定を必要とせず，何らかの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，ロボットの活躍する場面も更に拡大することが期待されている．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扱っている．

1.2　機械学習と強化学習

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ような活動を，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．大きく分けると教師あり学習，教師なし学習，強化学習の 3 種類である．

教師あり学習とはクラスラベルや関数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．付随情報が無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連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．

強化学習とは未知なる環境における適切な行動戦略を，経験を繰り返すことで獲得するタイプの学習アルゴリズムである[1][2]．強化学習の概念を図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び，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ー値を受け取る．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ている．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．その理由として自律エージェントや自律ロボットは実世界の複雑な環境で動作する．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ためである．そこで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．

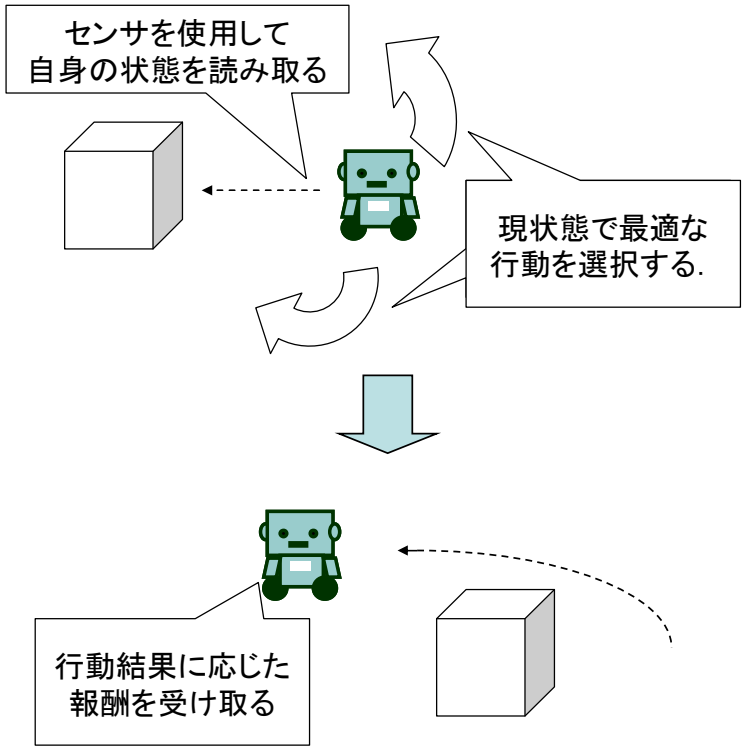


図 1：強化学習の概念

1.3 強化学習の問題点

強化学習を含む機械学習全般の問題の１つとして，学習結果を出すまでに時間がかかるという点がある．強化学習では，知識が無い状態から試行錯誤によって学習を行うため学習に時間がかかる．経験を繰り返すことで学習を行うため，エージェントが行動を行う環

境がより複雑で多様な状態になると環境の状態数が増加し学習に必要な時間が莫大なものとなる．例えば Watkins の Q-Learnings[3]では強化学習法は学習に大量の繰り返し計算が必要となり，学習結果の収束に時間がかかるという問題点があげられている．

また実ロボットに強化学習を適用させる際にも学習時間の問題が発生する．実ロボットでは，実際に機体を動かして行動するのに 1/10～1 秒程度，あるいはそれ以上の時間がかかる[4]．そのため実ロボットによる強化学習はコンピュータ上での強化学習よりさらに多大な時間が必要となる．

以上の 2 つの観点から強化学習の学習時間の増大を解決することが望まれる．

1.4 従来研究

強化学習の学習時間の削減に関する研究は主に以下の 2 つに分類される[2]．

（1）強化学習の試行回数の削減
（2）強化学習の計算時間の短縮

（1）の試行回数の削減とは行動の回数や目的達成するまでの試行回数を少なくすることである．具体例として報酬を過去の行動に伝播する適正度の履歴[5]，一度得た経験を何度も更新に用いるプランニング[6]，教示の導入[7]，マルチエージェントによる経験の共有[8][9]などがある．（2）の計算時間の短縮とは，１回の学習にかかる計算時間を削減することである．具体例として利得を近似して計算量を減らす Truncated Temporal Difference[10]，機構増を用いる TD（λ）の対数時間更新算法[11]，並列計算[12]などがある．

シミュレーション上での実験では，状態の取得や実際の行動に時間を要さない．そのため試行数が増えたとしてもあまり問題とならない．そのため（2）解決する研究が多数を占めている．一方実ロボットの場合では，状態の取得や行動の際にセンサからの読み取りや装置の稼働に時間を要する．そのため学習に必要な試行数が増えると，莫大な時間を要することになる．そのため実ロボットに強化学習を適用することを目的とする場合には（1）に関する研究が多くなる．

（1）の手法の 1 つに，ある 1 つのタスクを学習させる際に複数台の強化学習エージェントを同時に学習させる手法がある．複数台のエージェントが同時に学習を行い，経験を共有，または定期的に合成することにより，短時間で学習を収束させる方法が提案されている[8][9][13][14][15]．これらの手法を用いることにより，1 台のエージェントで単独学習を行うよりも短時間で学習を収束させることが可能となる．しかしこれら多くの手法では，学習中は全てのエージェントが共有，合成した情報に基づいた行動戦略を行う．そのためエージェント間の選択する行動が重複することによる学習効率の悪化や，探索の不足による全エージェントの局所解での収束などの問題が発生する場合がある．また実ロボットでマルチエージェント強化学習を行う場合，他のロボットとの情報のやりとりが問題となる[16]．

特に学習速度の向上を目的とした場合では複数台のエージェントによる学習情報の共有，または合成が行われる．そのためロボット間で通信する必要がある．このとき学習に必要となる情報を常に送信することになるため，通信が過大になりやすい．そのため通信の負荷やバンド幅などの面で問題が発生する．また複数のエージェントが同一環境で学習を行う場合，他のエージェントとの協調動作を行うことになる．協調動作を行うためにはそのための機構を各ロボットに搭載することになる．実ロボットによるマルチエージェント強化学習を導入するには，これらの問題を解決することが望まれる．

1.5　問題解決へのアプローチ

複数台の強化学習エージェントの同時学習の問題点は大きく分けて 2 種類ある．1 つはエージェント間の学習領域の重複，もう 1 つは実ロボットに適用した際のロボット間の通信である．前者の問題を解決する方法の 1 つとして各エージェントの学習領域が重複しないシステムを構築することが考えられる．後者の問題を解決する方法の 1 つとしてエージェント間の通信付加を下げるために 1 つのコンピュータ内，あるいは 1 つのシステム内でマルチエージェントを構成する方法が考えられる．

そこで本研究では単体のロボットにマルチエージェントを構成し，学習する領域を分割する方法を考える．マルチエージェントを 1 体のロボット内に構成することでエージェント間の通信が容易となる．また学習する領域を分割することで各エージェントの行動選択の重複を避けることができる．この提案手法に関する説明は第 3 章で詳しく説明する．

1.6　本研究の目的

強化学習の問題点の 1 つに「学習結果を出すまでに多大な時間を必要とする.」という問題点が存在する．そこで本研究ではこの問題を解決することを目的とする．

この問題を解決するに当たって，本研究ではマルチエージェント強化学習による手法に注目する．マルチエージェント法による強化学習を用いて学習速度を向上させる．しかしマルチエージェント強化学習を用いた手法にもいくつかの問題点が存在する．そこで本研究では 1 体のロボットにマルチエージェントを構成することでこれらの問題を解決する手法を提案する．

1.7　本論文の構成

第 1 章ではロボットの制御の背景から現在研究されている機械学習と強化学習について説明した．また強化学習の問題点を上げ，その問題点を解決するための従来研究を上げた．そして従来研究の問題点から目的達成のアプローチを示し，本研究の目的を述べた．

第 2 章では強化学習の基本的な枠組みと，強化学習の学習手法とその代表的な手法である Q-learning，また強化学習での学習の対象となる状態行動対について説明する．

第 3 章ではマルチエージェントシステムの特徴とロボットの行動の原理について説明す

る．そしてこの 2 つの要素に注目しマルチエージェントシステムを導入した提案手法について説明する．

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．そして実験結果からその考察を述べる．

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　強化学習

本章では，強化学習の概要について述べる．強化学習には学習手法と行動選択手法が存在するため各項目について説明する．特に強化学習の学習手法の中でも本研究で用いた Q-learning のアルゴリズムについて詳しく説明する．また強化学習の学習対象となる状態行動対について状態行動対を空間で表現した状態行動空間を使用して詳しく説明する．また行動選択手法についても説明する．

2.1　強化学習の枠組み

強化学習とは未知なる環境における適切な行動戦略を獲得するタイプの学習アルゴリズムである[1] [2]．強化学習の概要を図２に示す．強化学習では学習する主体をエージェント，学習システムにとっての外部からの情報を状態とする．状態とは環境からの感覚入力やエージェントの内部状態，あるいはそれらの組み合わせとなる．エージェントの目標をタスクと呼ぶ．強化学習では，エージェントは各時間ステップにおいて観測した情報を元に行動を決定する．その後実際に行った行動に対して環境から報酬と呼ばれるスカラー値を受け取る．報酬の大きさはタスクを表す状態に設定され，エージェントがタスク状態に達すれば，エージェントにより高い報酬が与えられる．強化学習の目的は最終的に得られる報酬の総数を最大化することである．エージェントは報酬の総数を最大化するために各状態における最適な行動を学習して導く．

強化学習ではエージェントは価値関数という関数を所持している．価値関数とはそれぞれの行動に対して，その行動が選ばれた場合の報酬の期待値を表す関数である．価値関数は状態と行動，そして評価の見積もり値で表される．エージェントは受け取った報酬を元に価値関数を計算し更新する．評価の見積もりとは各時間ステップにおける状態に対する行動の評価値を表している．したがって見積もり値が高ければその状態に対する行動が良いとされ，タスクを達成するために優先すべき行動と判断できる．この価値関数の計算方法は学習手法によって異なる．

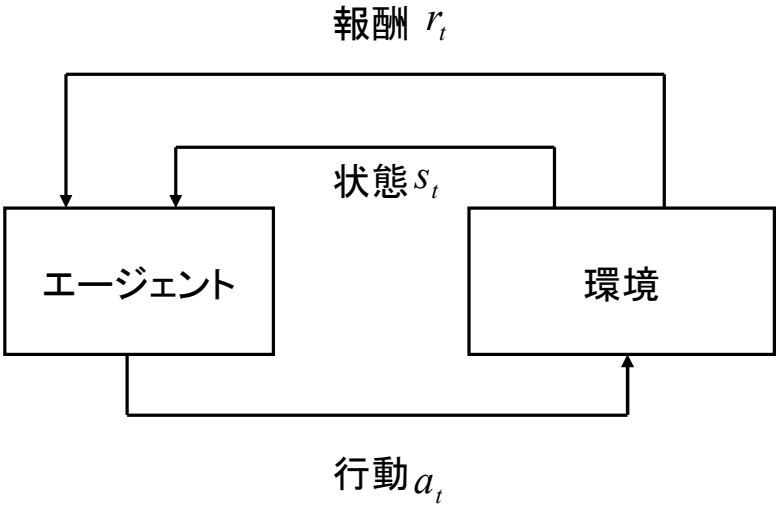


図 2：強化学習の概要

2.2 学習手法

強化学習には様々な学習手法が存在する．学習手法によって価値関数の計算の仕方が異なる．具体的な学習手法としてはTD(λ)法，Q-learning法，分類子アルゴリズムなどが存在する[17]．本研究では強化学習の学習手法としてQ-learningを使用する．

Q-learning では価値関数のことを行動価値関数という．行動価値関数とは状態と行動の組に対する評価を見積もる関数であり，それぞれの状態と行動の組の評価値をQ値と呼ぶ．Q-learning では時刻tで観測された状態s_tにおいてエージェントが行動a_tを実行し，時刻$t+1$で状態s_{t+1}に推移し，報酬r_{t+1}を得たとする．このとき状態s_tの行動価値関数$Q(s_t,a_t)$は式（1）を用いて更新される．

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

ここでαは学習率，γは割引率と呼ばれるパラメータであり，$0 < \alpha \leq 1$，$0 < \gamma \leq 1$を満たす実数である．

2.3 行動選択手法

強化学習では価値関数を元に行動を選択するが，その選択の手法は様々な方法が存在する．Q-learning などの強化学習法では，行動選択のアルゴリズムとしてε-greedy法やBoltzmann分布を用いた選択法が良く使われている．ε-greedy法では，状態sにおいて確率ε $(0 \leq \varepsilon \leq 1)$の確率でランダムな行動，確率$1-\varepsilon$でQ値が最大と成る行動を選択する．本研究ではε-greedy法を用いる．理由としてはε-greedy法はシンプルな手法で強化学習でも良く用いられる手法であり，ランダムに探索行動を行うため十分な探索を行うためには何度も試行を重ねなければ成らないため本研究で扱う問題点が特に現れるためである．

2.4　状態行動対

状態行動対とはQ-learningにおける評価の見積もりとなる行動価値関数を構成する状態と行動の組みのことをいう．強化学習では価値関数を計算し更新することでタスクを達成できる行動を獲得する．そのため価値関数を構成する要素が多いほど，学習を行う領域が大きくなる．Q-learning では状態行動対が学習領域となる．また価値関数を構成する要素は状態と行動になる．そのため状態値や行動数が増加することで状態行動対が多くなれば学習する領域も広くなり，学習が収束するまでに時間がかかるということになる．

この問題を状態行動空間の観点から示す．状態行動空間とは状態行動対を表空間上で表したものである．状態行動空間の例を図３で示す．状態行動対は Q-learning の行動価値関数を構成する要素である．そのため状態行動空間は状態と行動が軸となり構成される．状態軸はエージェントの現在の状態を示す値，行動軸はエージェントが出力する各行動となる．図３の四角で囲まれた部分が 1 回の行動で学習する状態行動対となる．最適な行動を得るためには全ての状態行動対を経験し学習することが望まれる．これは状態行動空間では全ての領域を経験し学習するということである．したがって状態行動空間が大きいほど学習に時間がかかるということである．

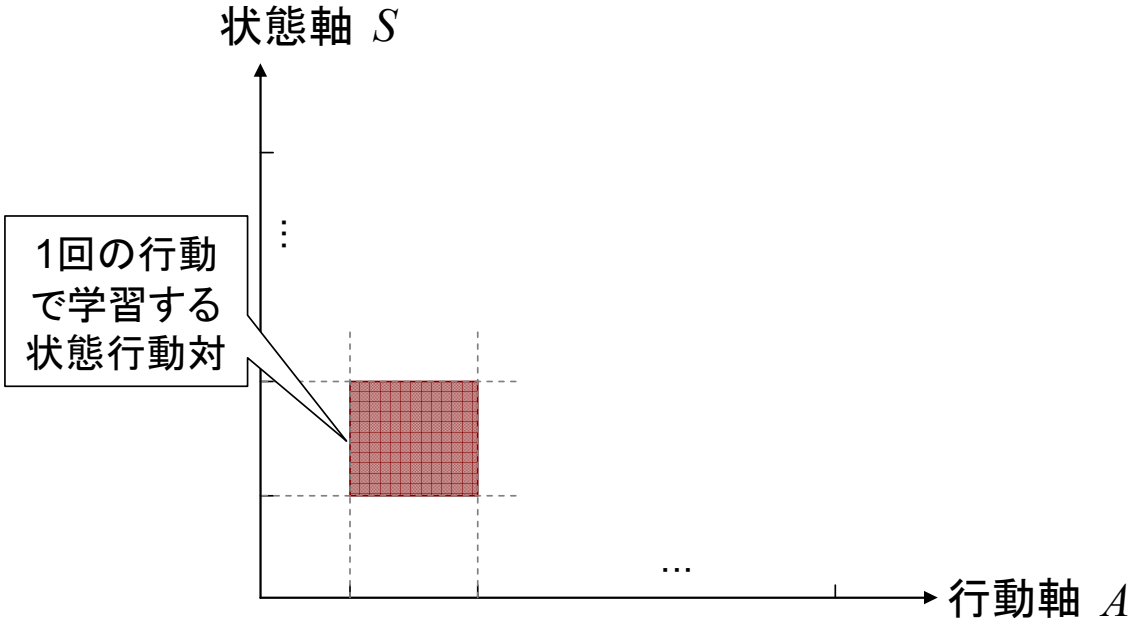


図 3：状態行動空間の例

図 4 では状態行動空間の行動軸が増加した例を示す．強化学習では環境に対して十分な経験を得るまでに何度も試行を重ねるため，学習に時間がかかる．搭載するセンサの数の

増加やロボットの内部状態を構成する要素の増加よる環境数の増加，ロボットの実行可能な行動の増加による行動数の増加が発生する．この環境数の増加や行動数の増加は状態行動空間では状態軸や行動軸の数が増加することで表される．エージェントの実行可能な行動数が増えた場合，行動の分割数が増加する．そのため学習する領域が増大し学習が収束するまでの時間が増大する．

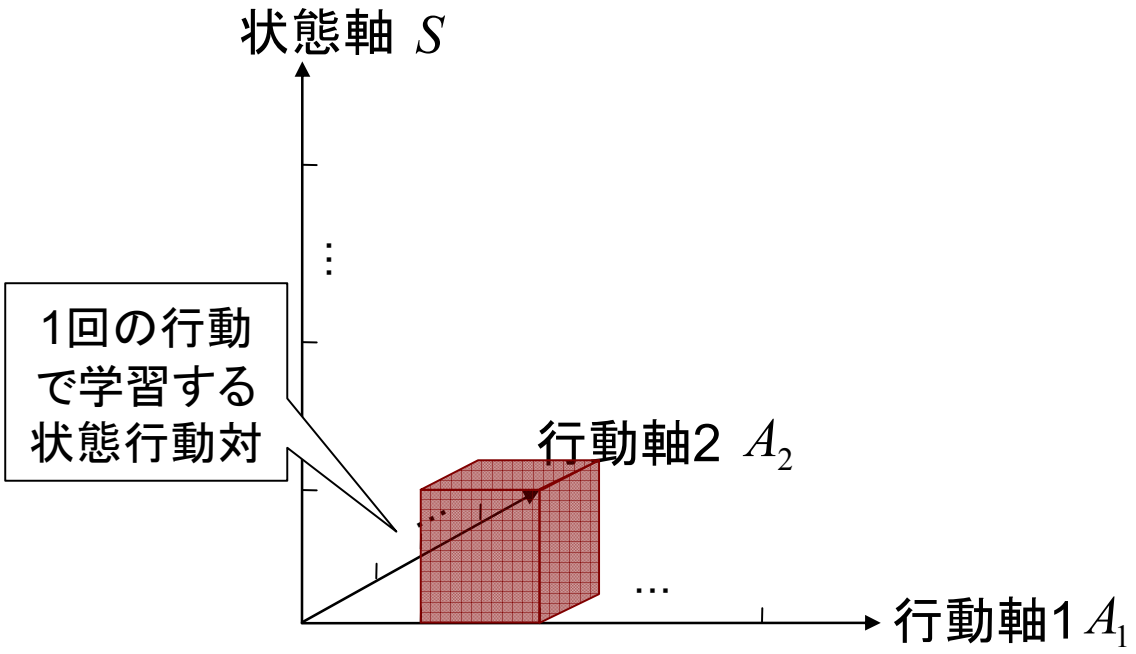


図 4：エージェントの実行可能な行動が増加した際の状態行動空間の例

第 3 章　提案手法

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．まず提案手法で適用するマルチエージェントシステムについてと，1 体のロボットでマルチエージェントを適応するために注目したロボットの行動に関する説明を行う．そしてこの 2 つの要素から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．

3.1　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．マルチエージェントシステムの概要を図 5 で示す．それぞれのエージェントは自身の環境を知覚して，自分の目標を達成するように行動をとる．マルチエージェントシステムの大きな特徴は，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である．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を行うことが重要となる．

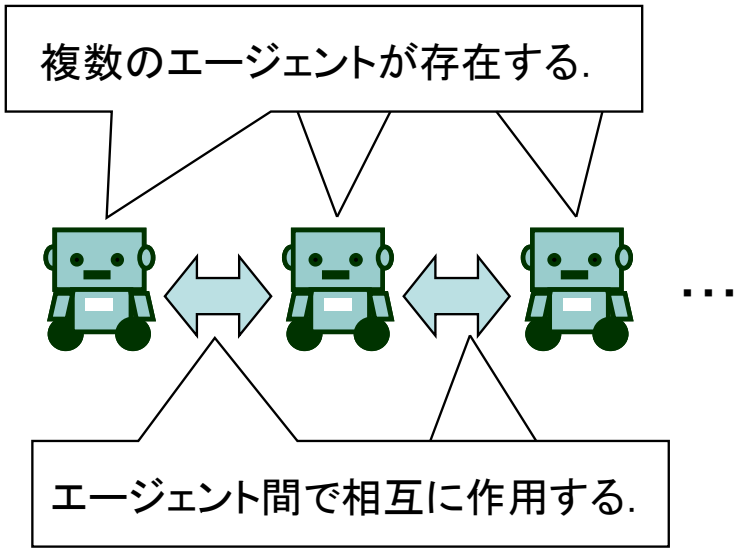


図 5：マルチエージェントシステムの概要

3.2　ロボットの行動

「ロボットが行動する」というのは，ロボットに搭載されている駆動装置を動かすことである．このロボットに搭載されている駆動装置のことをアクチュエータと呼ぶ．ロボットが他の機械と大きく異なる点は，実環境に対して何らかの影響を与えられる点である．

そのためアクチュエータはロボットにとって非常に重要となる．ロボットに搭載するアクチュエータの種類や数によってロボットが実行可能な行動は変化する．アクチュエータの種類にもよるが，搭載するアクチュエータの数が多ければロボットの実行可能な行動は多くなる．またロボットの行動は 1 つのアクチュエータの動作によって構成されているとは限らない．複数のアクチュエータを同時に稼動し協調動作させる事により 1 つの行動を構成する場合もある．したがってロボットの行動は搭載されているアクチュエータの協調動作によって構成されていることになる．複数のアクチュエータの動作の協調動作の例を図６に示す．この例ではアクチュエータ１・２・３のそれぞれの動作 A1・A2・Ａ３が協調動作によってロボット全体の行動 A が生成されている．

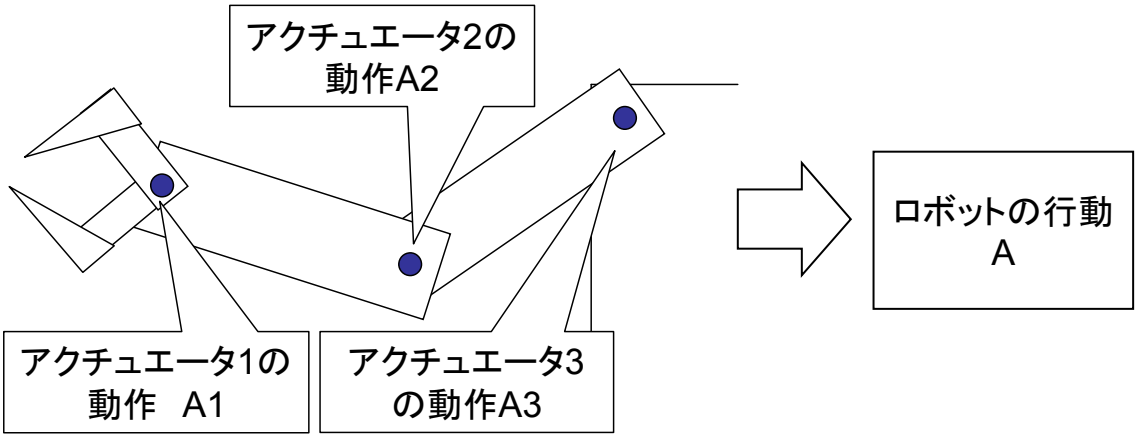


図 6：各アクチュエータによる動作の協調動作の例

3.3　提案手法の概要

ロボットの行動は各アクチュエータの動作の協調動作によって決定される．各アクチュエータの動作が決定すればロボットの行動は一意に決まる．そこで本研究ではロボットに搭載されているアクチュエータ毎にエージェントを設定し，マルチエージェント強化学習を構成するシステムを提案する．アクチュエータ毎によるマルチエージェント強化学習の概要を図７で示す．強化学習における行動は学習の対象でありシステムの出力である．エージェントをアクチュエータ毎に設定した場合，出力されるものはアクチュエータの動作である．したがって各エージェントは現在の状態における最適な自身のアクチュエータの動作を学習することになる．またマルチエージェントシステムの特徴の 1 つに各エージェントは他のエージェントと協調して動作を行う点がある．各エージェントが他のエージェントと協調して動作を決定すればロボット全体の挙動を生成することができる．ロボット全体の挙動を生成することができれば，ロボット全体を 1 エージェントとした場合とアクチュエータ毎にエージェントを設定した場合でロボット全体の挙動は変わらない．また強化学習では環境から受け取る報酬を元に学習を行う．各エージェントはそれぞれ総報酬の値が最も高くなる動作を獲得することを目標とする．そのためロボット全体を 1 エージェ

ントとした場合とアクチュエータ毎にエージェントを設定した場合で，ロボット全体の挙動が変わらなければ学習が収束した際の受け取る報酬の値も同じとなる．そのためアクチュエータ毎にエージェントを設定しても，ロボット全体はタスク達成が期待できる．

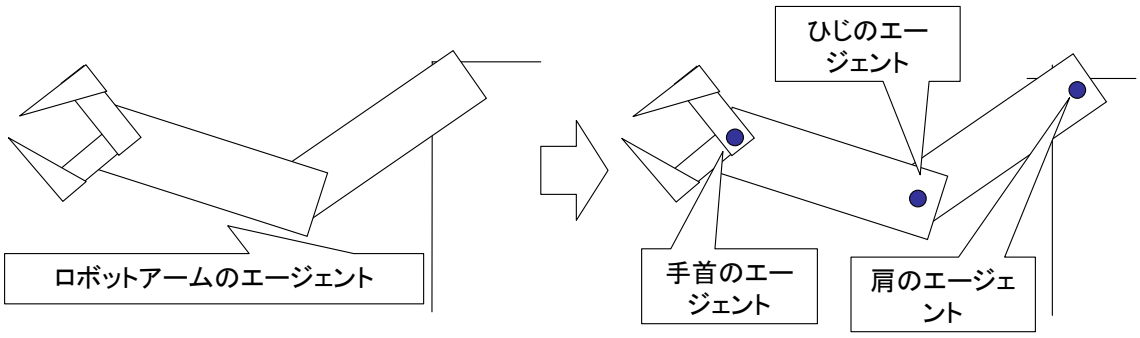


図 7：提案手法の概要

またアクチュエータ毎にエージェントを設定した際の状態行動空間の変化を図８に示す．エージェントの行動をアクチュエータの動作と考えると，状態行動空間の行動軸は各アクチュエータの動作となる．そのため各行動軸はそれぞれ 1 つのアクチュエータの動作を表す．図８は 2 つのアクチュエータが搭載されている場合の例である．アクチュエータ毎にエージェントを設定した場合，状態行動空間はエージェント毎に分割される．各エージェントの状態行動対は状態軸については変化せず，行動軸は各アクチュエータの動作軸となる．マルチエージェントシステムの場合，各エージェントが同時に学習を行うことができる．そのため各エージェントの行動価値関数を更新することができる．また各エージェントの行動はアクチュエータの動作となる．各アクチュエータの動作が重複することは無いため各エージェントの出力は重複しないことになる．したがって各エージェントの状態行動対は重複しなくなるため，従来手法に存在した各エージェントの行動の重複問題は発生しなくなる．またロボットの行動は各アクチュエータの動作の組み合わせで決定される．そのため各アクチュエータの動作数はロボット全体の行動数より少なくなる．そのため状態数が変化しなければ各エージェントの状態行動対はロボット全体の状態行動対より大きくなることはない．

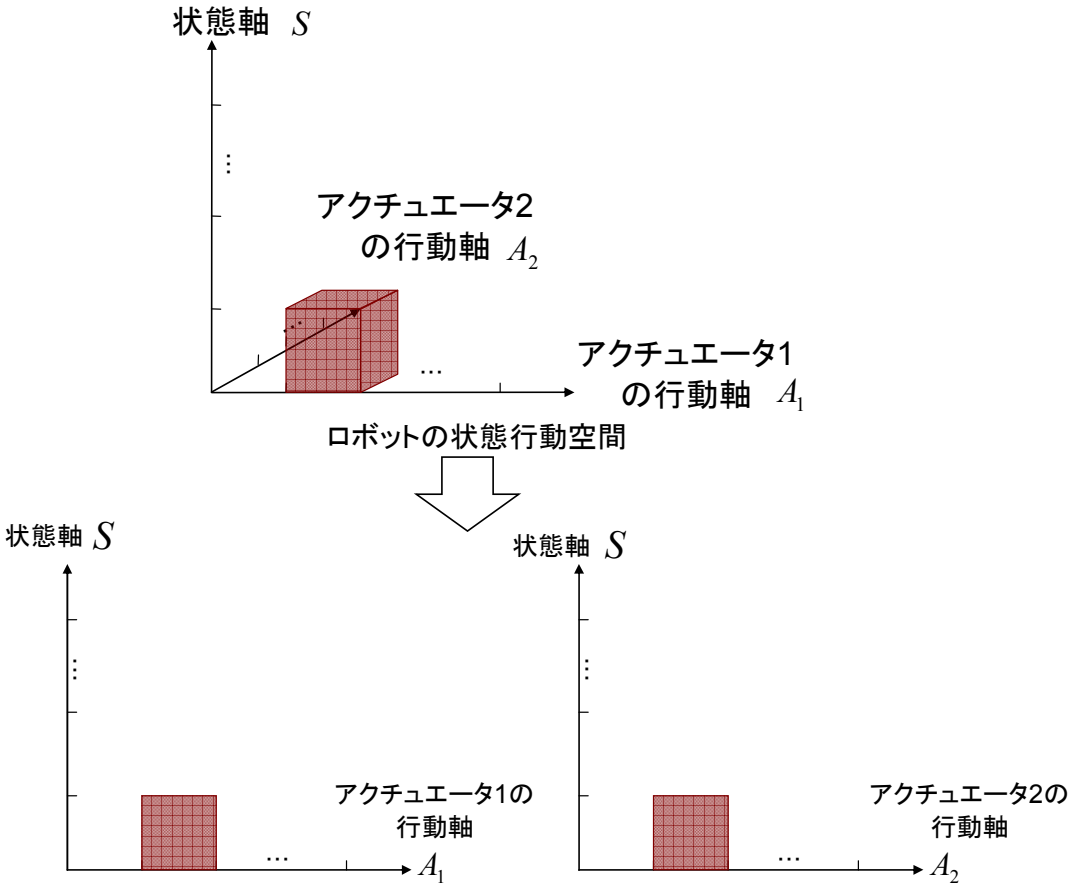


図 8：提案手法を導入した際の状態行動空間の変化

3.4　提案手法の構成

提案手法の概要を図 9 に示す．ロボット全体の状態を取得した後，各エージェントが取得した状態値を元に動作を選択する．全エージェントが動作を決定した後実際にロボットが行動する．行動後，ロボットは環境から遷移に応じた報酬を受け取る．その報酬を元に各エージェントが自身の選択した動作に対して学習する．この流れを強化学習の 1 ステップとする．

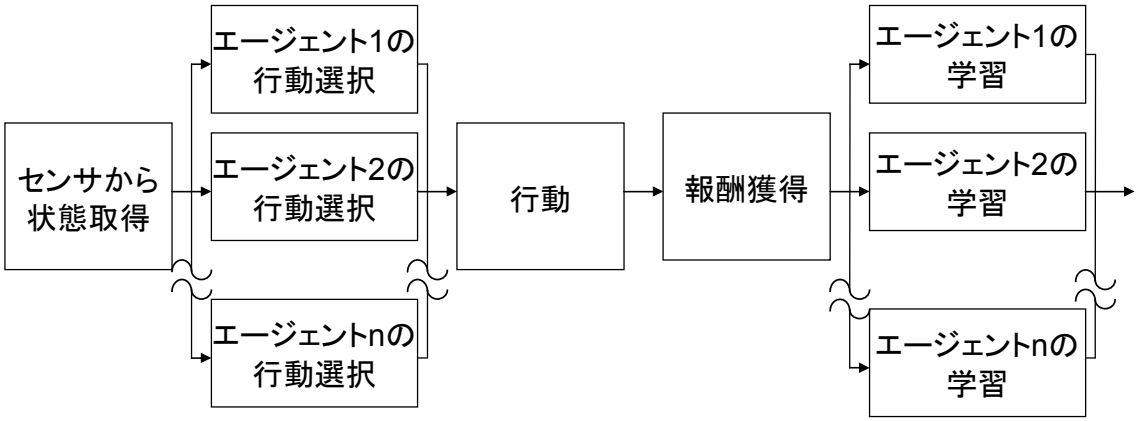


図 9：提案手法の概要

マルチエージェントシステムにより強化学習を正しく行うためには，エージェント同士が協調動作する必要がある．そのため本研究では 2 つの方法を用いてエージェント同士の協調動作を行う．

（1）各エージェントは自身を含む全てのエージェントの状態を認識する．
（2）全アクチュエータの動作は同期させ一斉に処理する．

（1）の「全エージェントの状態を認識する」とは，行動選択と学習の際に取得する状態に他のエージェントの状態を取得することである．全エージェントの状態認識の概要を図 10 に示す．ロボットに強化学習を適用させる際には，状態は外部状態と内部状態の 2 種類が存在することになる．外部状態はセンサから取得する環境を数値情報で表したものである．内部状態はエージェント自身の状態を表す．ロボット全体を 1 エージェントとした際の内部状態はロボット自身の状態を表す．各アクチュエータの状態はロボット自身の状態を表す情報の 1 つである．一方，各アクチュエータを 1 エージェントした際には内部状態は自身のアクチュエータの状態と表す．しかし実際に行動し報酬を受け取るのは 1 体のロボットである．そのため各エージェントは自身のアクチュエータの状態とロボットが置かれている環境の状態だけではロボット全体の状態を正確に把握できない．ロボット全体の状態を正確に把握できなければロボットがタスクを達成しているかを判断できず正確な学習が行えない．そのため各エージェントはロボット全体の状態を詳細に認識する必要がある．したがって各エージェントは自身を含む全てのエージェントの状態を認識できるようにする．

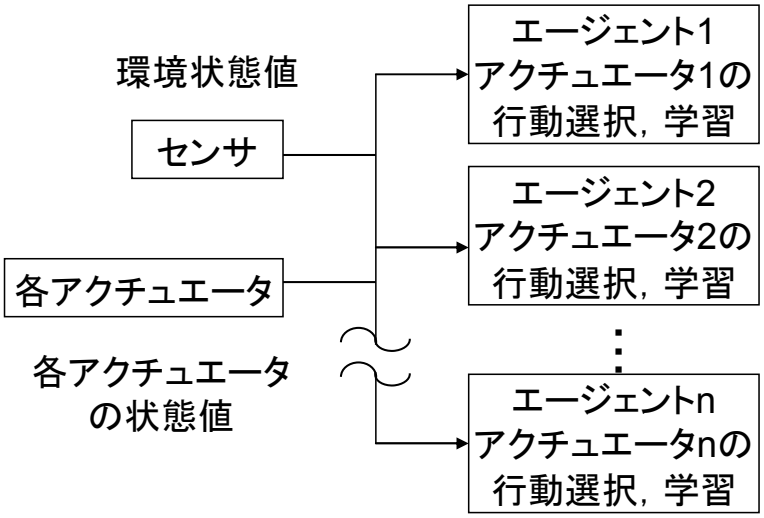


図 10：全アクチュエータの状態認識の概要図

（2）の「全アクチュエータの動作は同期させる」とは各エージェントが行動選択を完了し実行する際に，それぞれが独立で動作するのではなく全エージェントが一斉に処理を行うことである．各エージェントは自身の行動を決定した後，行動を決定した信号を送信する．全エージェントの行動が決定した後，全てのアクチュエータを一斉に行動させることで同期させる．各アクチュエータ動作の同期の概要を図 11 に示す．アクチュエータの動作はロボットの外部状態，内部状態のどちらにも何らかの影響を与える．それぞれのエージェントが独立に動作した場合，あるエージェントが行動選択や学習を行っている最中に他のエージェントが動作してロボットの状態が変化する場合がある．このような状況が発生すると，同じ状態で同じ行動を選択しても動作後や学習後の状態が一意に定まらない．状態が一意に定まらないと報酬の値も変化してしまうため各エージェントの学習が正しく行えない．そのため各エージェントの動作を同期させ一斉に処理することで，行動選択や学習の際に各エージェントが取得する状態値を一意に定める．

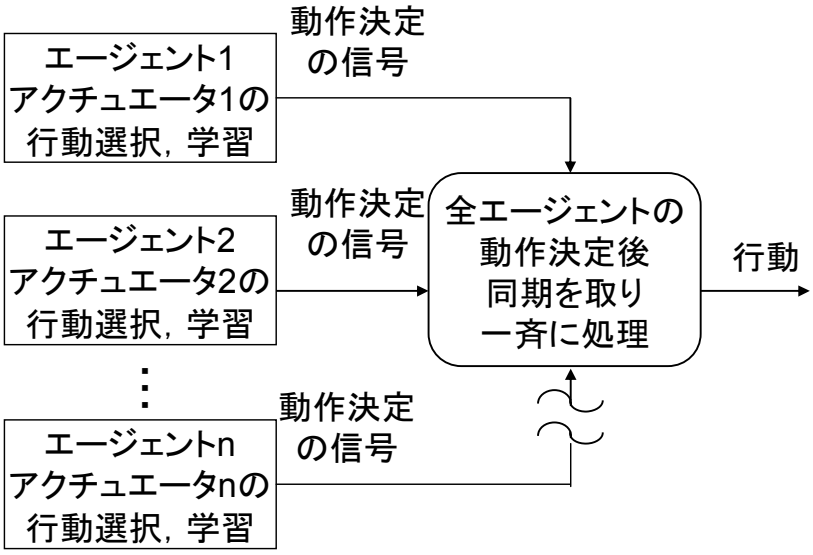


図 11：各アクチュエータ動作の同期の概要

この 2 つの方法を採用するためには各エージェントを総括するシステムを構築する必要がある．各エージェントを統括する方法としては，1 つのコンピュータでマルチエージェントシステムを構築し処理を行う方法，各エージェントを統括するコンピュータをロボット内に加える方法，エージェント間で通信し動作決定の信号を交換する方法の 3 種類が考えられる．1 つ目の方法では，各エージェントの処理を逐次的に行うシステムとなる．2 つ目の場合では統括するコンピュータが各エージェントに現在の状態値を送り，全エージェントの動作決定の信号を確認し実際にアクチュエータを動作させる命令を送るシステムとなる．3 つ目の場合では各エージェントが動作決定の信号を送り全エージェントの信号が送られた時点で実際に行動するシステムとなる．いずれの手法を用いた場合でも提案手法の基本的なシステムは同様となるように構築する．

3.5　システム処理の詳細

提案手法のシステム構成を図 12 で示す．　各エージェントが行動選択，学習を行う際に必要となる状態値は，すべて共通とし同じ時刻での情報を送る．送る情報は現在のロボットの環境状態値と現在の各アクチュエータの状態値となる．各エージェントは行動を決定した後には，システム全体で同期させることでロボット全体の挙動とする．

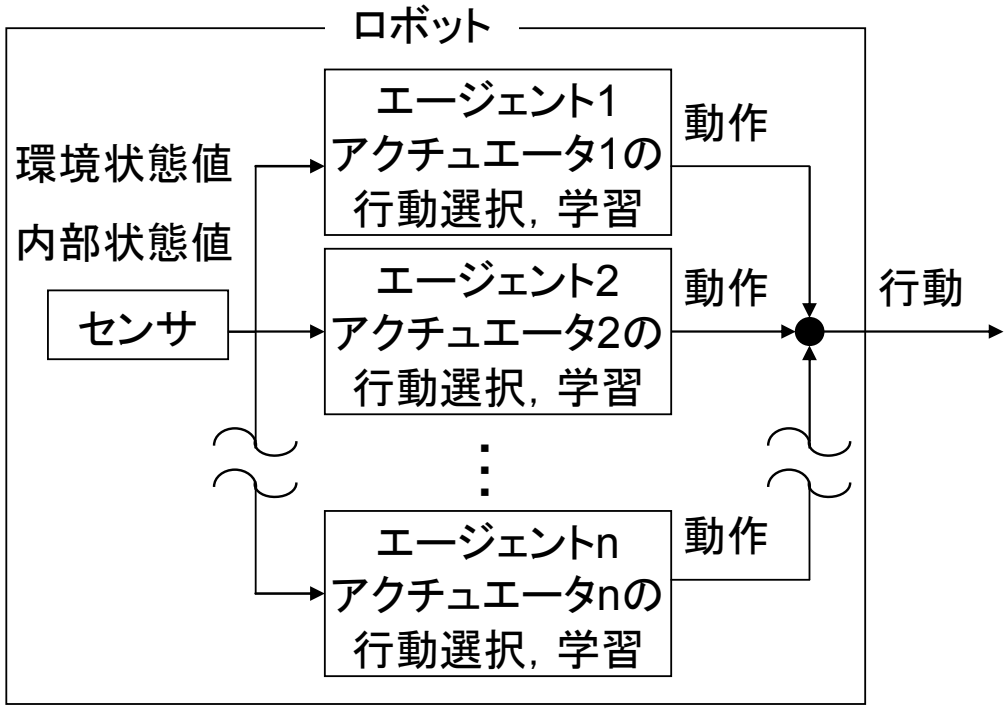


図 12：提案手法のシステム構成

提案手法のシステムの処理の流れを図 13 に示す．まずセンサからロボットの現在の環境状態を取得する．この取得した環境状態値と全アクチュエータの状態値を各エージェントに送る．各エージェントは送られてきた情報と現在の行動価値関数を元に，現時点での最適な動作を選択する．各エージェントは自身の行動を決定した後，実際にアクチュエータは動かさずに行動決定の信号を送る．全エージェントが動作決定の信号を送った後，同期をとり全アクチュエータの動作を一斉に実行する．ロボットが行動を終えた後，ロボットは環境から遷移状態に応じた報酬を受け取る．各エージェントは受け取った報酬を元に自身の選択した動作について学習を行う．この一連の流れを 1 ステップとして学習を進める．

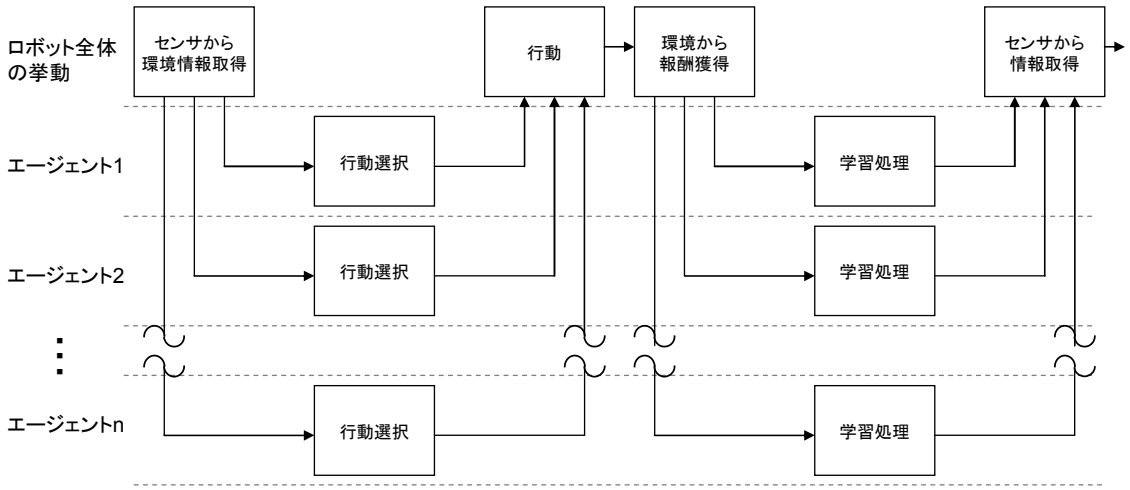


図 13：提案手法のシステム処理の流れ

第 4 章　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の説明をする．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有用性を確かめるために，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の学習が収束し，従来手法と同じ行動を獲得し同じ値の報酬を獲得している点と，本研究の目的である学習収束までの時間の短縮を達成している点の 2 点である．本研究での従来手法とは，1 ロボット 1 エージェントによる強化学習のことである．

4.2　実験概要

本節では本研究で行う実験について説明する．実験を設定する上の注意点として，正確な比較を行うために提案・従来手法を除きのタスクや想定する実験環境，ロボット全体が出力する挙動は従来手法と提案手法で共通とする点がある．特にロボット全体の挙動は設定段階で注意する必要がある．従来手法ではロボット全体の挙動で設定されるが，提案手法では各アクチュエータの動作を設定することになる．そのため各アクチュエータの動作の組み合わせとロボット全体の挙動が一致するようにタスクを構成する．また提案手法は各エージェントが協調動作を行うことでタスクを達成する．この内容を検証するために実験で行うタスクでは 2 つのアクチュエータの協調動作が必要となるタスクを設定する．

4.2.1 タスク設定

本研究では台車ロボットの荷物運搬タスクを選択し，実験を行った．荷物運搬タスクの概要を図 14 に示す．台車ロボットには傾斜角度を変化させることができるテーブルが搭載されている．運搬する荷物はテーブルの上に乗せると仮定する．台車ロボットの目的は，任意の目的地に荷物を運搬することである．台車ロボットは加速度を変化させることで目的地にたどり着く．しかしテーブルの上にものを乗せた際には，ものを落としてはいけない．そのため台車ロボットは台車の加速度とテーブルの角度を調節することで，テーブルの上のものを落とさない状態を保たなくてはならない．したがって台車ロボットの目的は「テーブルの角度をものが置かれた際に，ものを落とさないためのテーブルの角度を保ちつつ，目的地に到達する.」こととなる．

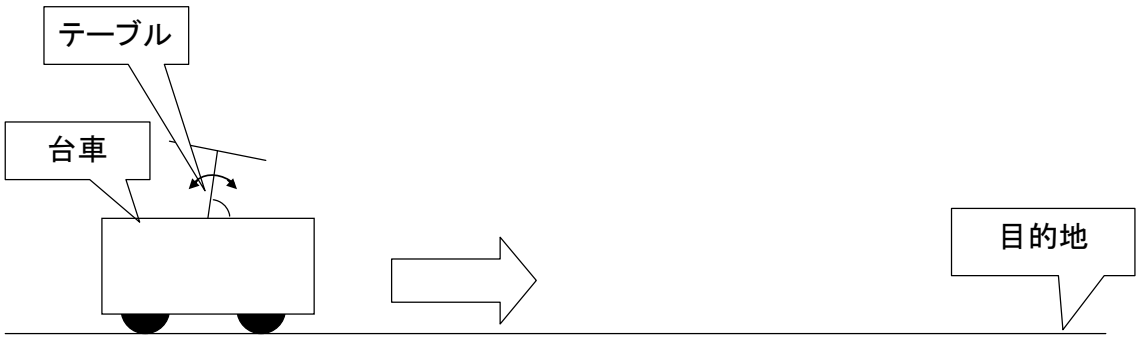


図 14：荷物運搬タスクの概要

台車ロボットには台車を動かすアクチュエータとテーブルの角度を変化させるアクチュエータが搭載されている．台車ロボットの詳細図を図 15 に示す．台車の前後の向きはスタート地点から見た目的地の方向が前方とする．加速度の方向は台車の前方方向を正とする．角度の方向は，加速度と同じ方向とし．地面から垂直に立っている状態を 90°，台車の前方に倒れている時を 0°，後方に倒れている時を 180°とする．

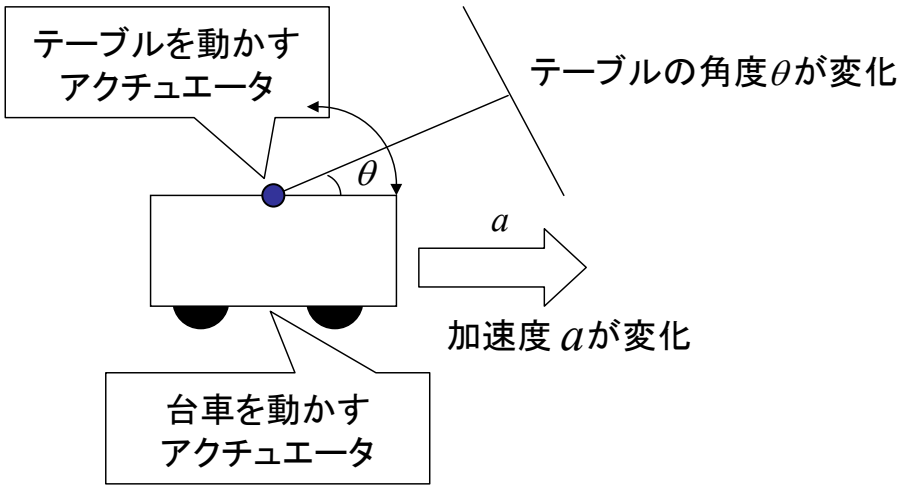


図 15：実験で想定する台車ロボット

台車のアクチュエータに設定する行動は，加速度の変化である．加速度の変化率は，テーブルの角度の変化率とつり合うように設定する．台車の行動は，加速度 a の値を $+0.5(m/s^2)$，$\pm 0(m/s^2)$，$-0.5(m/s^2)$ の 3 種類とする．ただし台車の速度を $v(m/s)$ とした時に，台車の制限速度 $-2 \le v \le 2$ を超える加速度を出力することはできない．台車の加速度 a の範囲は $-1 \le a \le 1$ とする．したがって台車の加速度は 5 状態存在する．

テーブルのアクチュエータに設定する行動は．テーブルの角度の変化である．角度の変化率は，加速度の変化率とつり合うように設定する．テーブルの行動は，テーブルの角度 θ の値を $+3°$，$\pm 0°$，$-3°$ の 3 種類とする．テーブルの角度 θ の範囲は $78 \le \theta \le 102$ とする．

したがってテーブルの角度は 9 状態存在する．実際のエージェントの行動は従来手法と提案手法でエージェントの数や設定方法が異なるので 4.2.2 と 4.2.3 でそれぞれについて説明する．

台車ロボットの認識する状態は台車の加速度a，テーブルの角度θ，台車の現在位置の 3 種類とする．台車ロボットに搭載する位置センサの状態値の設定方法を図 16 に示す．台車の現在位置の状態値$x_i(i=0,1,2\cdots11)$は台車の走行距離　を$X(m)$，台車のスタート地点を$X_0=0(m)$とした時に，$0\le X<100$の範囲では　$10(i-1)\le X<10i$の時$x_i=i$とする．また$X<0$の時は$x_0=0$，$100\le X$の時は$x_{11}=11$とする．したがって台車の現在位置は 12 状態存在する．これら 3 つの認識する状態値から，台車ロボットの認識する全状態数は$3\times9\times12=324$状態となる．

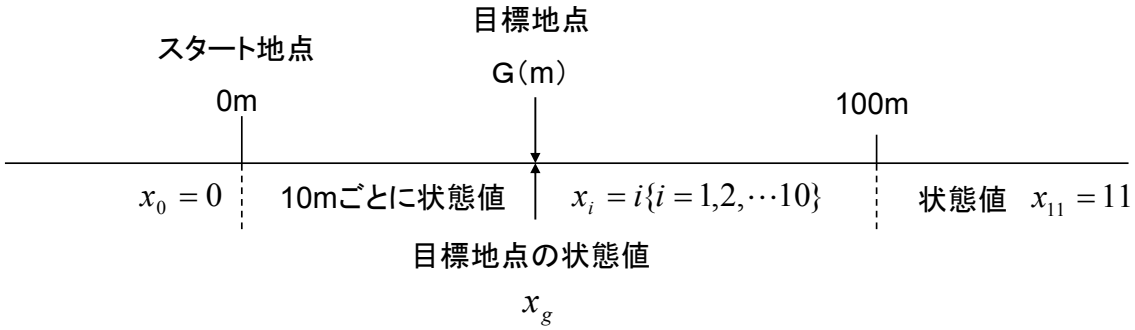


図 16：位置センサの状態値の設定方法

報酬rは式（2）で決定される．報酬はテーブルの角度がものを落とさないための角度を保っているほど高く、また台車の位置が目的地に近いほど高い報酬を得る．今回の実験はシミュレーションのため，テーブルの角度の最適な状態は，台車の加速度と重力によって発生する合力に対して垂直である状態とする．台車の現在位置の状態値をx，スタートから見た目的地の位置状態値をG，水平面から見たときのテーブルの角度をθ，水平面から見たときの合力をRとする。またw_1，w_2，w_3は係数である．

$$r=w_1(\theta-R)^2+w_2(G-x)^2+w_3 \qquad \cdots(2)$$

4. 2. 2 従来手法を用いたロボット

提案手法の比較対象となる従来手法とは 1 ロボットに対して 1 エージェントによる強化学習を適用した手法のことである．従来手法を用いた場合ではロボット全体で 1 つのエージェントとなるので，エージェントの行動はロボット全体の挙動となる．そのため従来手法を用いたエージェントは台車とテーブルの行動を決定する．1 ロボット 1 エージェントの

際のロボットの行動を図 17 に示す．台車の行動は 3 種類，テーブルの行動が 3 種類となるので，エージェントの行動は$3\times3=9$種類となる．

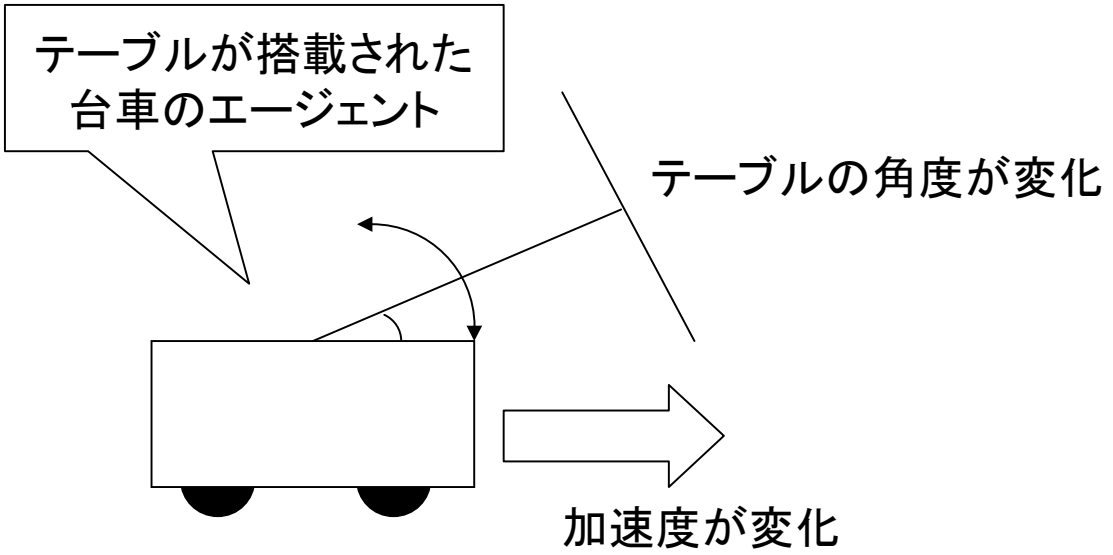


図 17：従来手法のエージェントと行動

4.2.3 提案手法を用いたロボット

本研究で提案する手法では，ロボットに搭載されている各アクチュエータにエージェントを設定する．本研究の実験で使用するロボットには台車を動かすアクチュエータとテーブルを動かすアクチュエータが搭載されている．したがって提案手法を用いる場合ではエージェントは台車のアクチュエータと，テーブルのアクチュエータの 2 つが設定される．各エージェントとその動作を図 18 に示す．台車アクチュエータに設定されるエージェントの動作は台車の加速度を変化させるもので 3 種類である．テーブルアクチュエータに設定されるエージェントの動作は，テーブルの角度を変化させるもので 3 種類である．この 2 つのアクチュエータの動作の協調動作によってロボット全体の挙動が生み出される．そのためロボット全体の挙動は$3\times3=9$種類となる．したがってロボット全体の挙動は従来手法を用いたロボットと一致する．各エージェントは出力以外のパラメータを共通とする．出力以外のパラメータとは行動選択や学習の際に取得する状態値，行動後に環境から受け取る報酬値である．

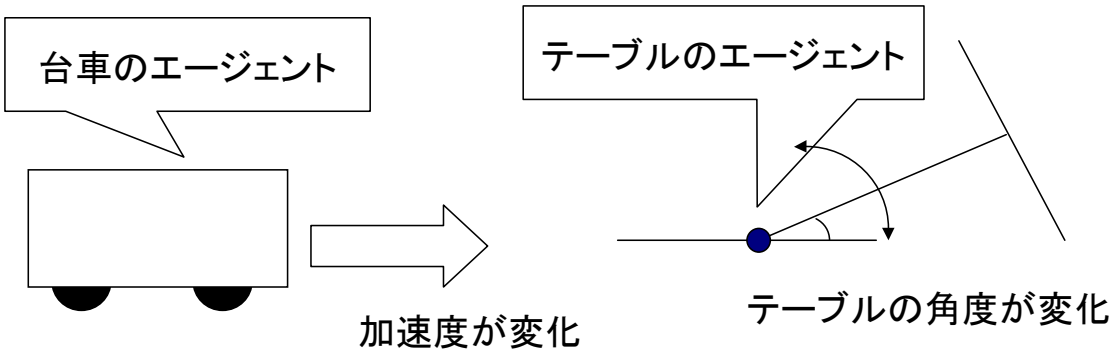


図 18：提案手法の各エージェント動作

4.2.4 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 1 に示す．台車ロボットの行動は 1 秒毎に行い，1000 回行動を 1 試行とする．目的地にたどり着いたとしても 1000 回行動するまでは行動し続ける．1000 回行動し終えた後，次の試行に移る．また今回の実験では台車の目標地点を距離センサの状態値 $x_G = 5.5$ と設定した．報酬式の係数は報酬の最大値が 0 となるように設定した．

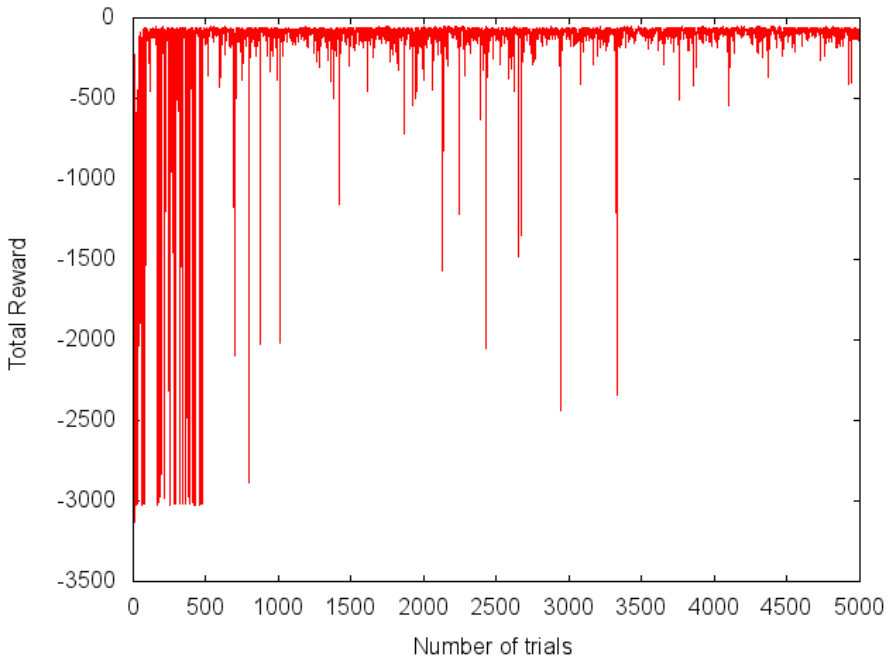
表 1：実験パラメータ

試行回数	5000（回）
1 試行の行動回数	1000（回）
行動の間隔	1 秒毎
学習を行うタイミング	1 行動毎
目的地の位置	状態値 $x_G = 5.5$（実際の距離 $50(m)$）
重力加速度 g	$9.8(m/s^2)$
行動学習手法	Q-learning
行動選択手法	ε -greedy 法
初期値	加速度 $a : 0(m/s^2)$ 角度 $\theta : 90(°)$ 速度 $v : 0(m/s)$ 位置の状態値 $x : 0$（実際の位置 $X : 0(m)$） 各 Q 値 $Q(s_t, a_t) : 0$
報酬式の係数	$w_1 : -10 \quad w_2 : -0.1 \quad w_3 : 0$
ε	0.05
ステップ・サイズ・パラメータ α	0.5
割引値 γ	0.5

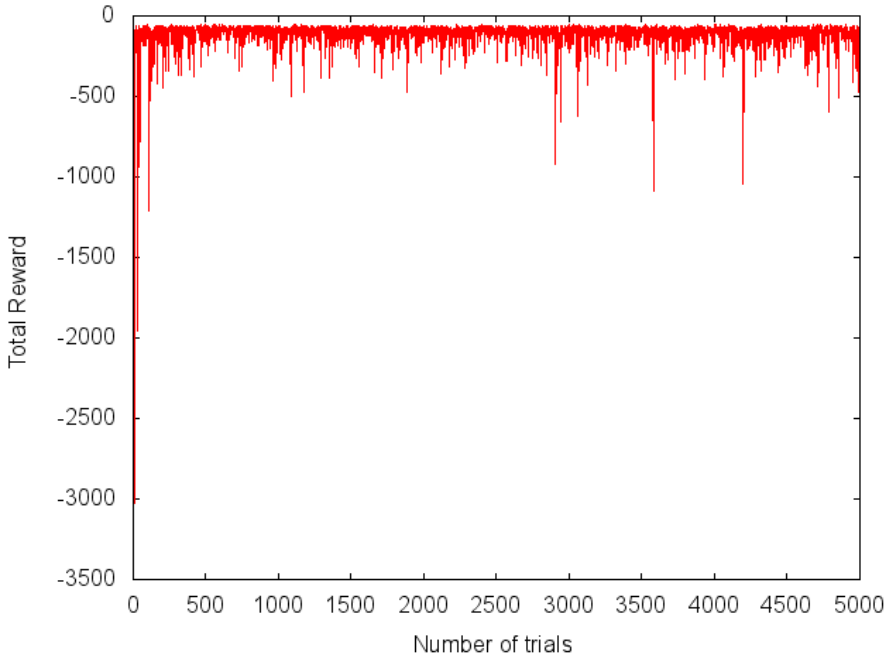
4.3　実験結果

本節では前節の設定で行った実験の結果を示し，その結果について解説する．提示する実験結果は，1 試行毎の総獲得報酬の推移，各試行の行動数と走行距離の変位，各試行の行動数とテーブルの加速度と重力の合力とのつりあいの変位，各試行の行動数と獲得報酬の変位を示す．それぞれの結果は従来手法を用いたロボットと提案手法を用いたロボットの結果を示し，結果について手法ごとに比較する．

初めに従来手法と提案手法それぞれの 1 試行毎の総獲得報酬の推移を図 19 示す．図 19 の横軸は試行数，縦軸は 1 試行の総獲得報酬である．今回の実験では報酬 r は $r \leq 0$ と設定している．そのため総獲得報酬が 0 に収束していれば学習が収束しているといえる．図 19 から従来手法，提案手法共に 0 付近で収束していることが分かる．したがって従来手法，提案手法共に学習が収束していることがいえる．収束するまでの試行数を比較すると従来手法を用いたロボットは約 500 試行目まで総報酬の値が低くなっているが，提案手法を用いたロボットは従来手法を用いたロボットよりも試行数が少ない段階で収束していることが分かる．この結果から提案手法は従来手法より少ない試行数で学習が収束しているといえる．また従来手法を用いたロボットは 2000 試行目以降で総獲得報酬の値が極端に低くなっている部分が何箇所か存在する．これは未探索の状態行動対が残っており一時的に迷い込みが発生しているためである．一方で提案手法を用いたロボットは何箇所か総獲得報酬の値が低くなっている部分が存在するが，従来手法を用いたロボットほど低い値ではない．また低くなっている部分の数も従来手法を用いたロボットより少ない．これは提案手法の方が学習収束までの試行数が少なくなるため，未探索の状態行動対になった際にも少ない試行数で適切な行動を獲得しているためと見られる．この結果から提案手法を用いたロボットは正しく学習を行い，学習が収束することが示された．これは提案手法の各エージェントの協調動作が適切に行われていることを示す．



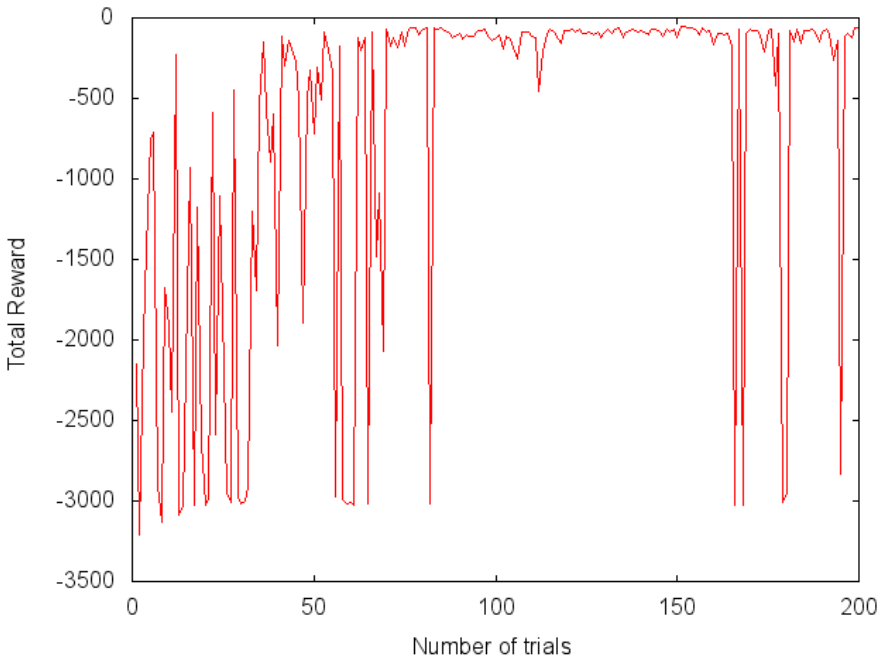
(a)：従来手法



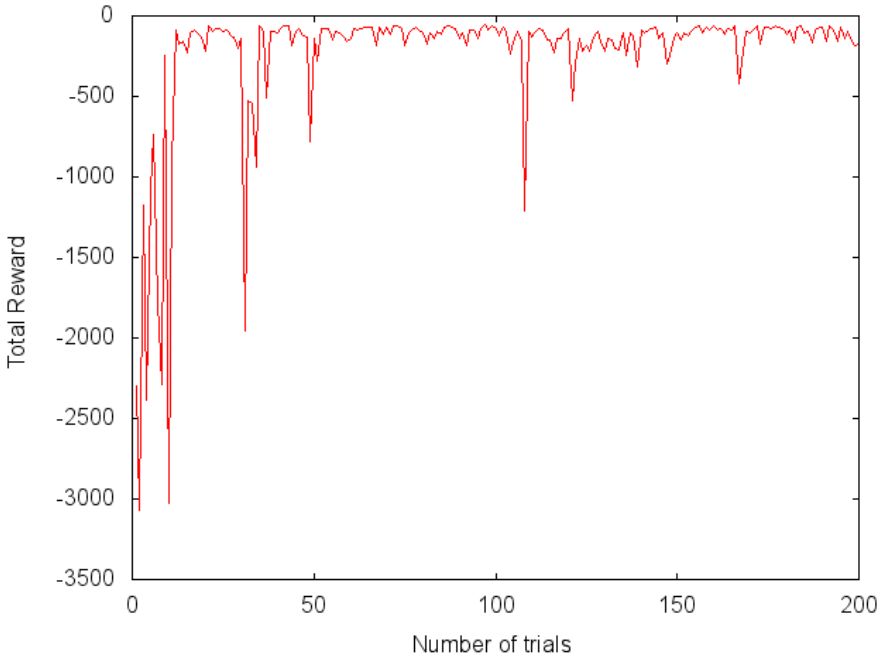
(b)：提案手法

図 19：各手法の 1 試行毎の総獲得報酬の推移

次に図 19 の試行数が 0 回目から 200 回目の範囲を拡大したグラフを図 20 に示す. 図 20 を見ると, 従来手法を用いたロボットは 50 回目から 100 回目の間で総獲得報酬の値が安定し始めている．また 150 回目以降に再び総報酬の値が低くなっている．これはロボットが未探索の状態行動対に突入したため再び学習を始めたためである．それに対して提案手法を用いたロボットでは 50 回目の前で総獲得報酬の値が安定し始めていることが分かる．この結果から提案手法を用いたロボットのほうが従来手法を用いたロボットより少ない試行数で学習が収束することが言える．これは提案手法の状態行動対のマルチエージェントによる分割が，学習収束の速度の向上に有効であることを示している.



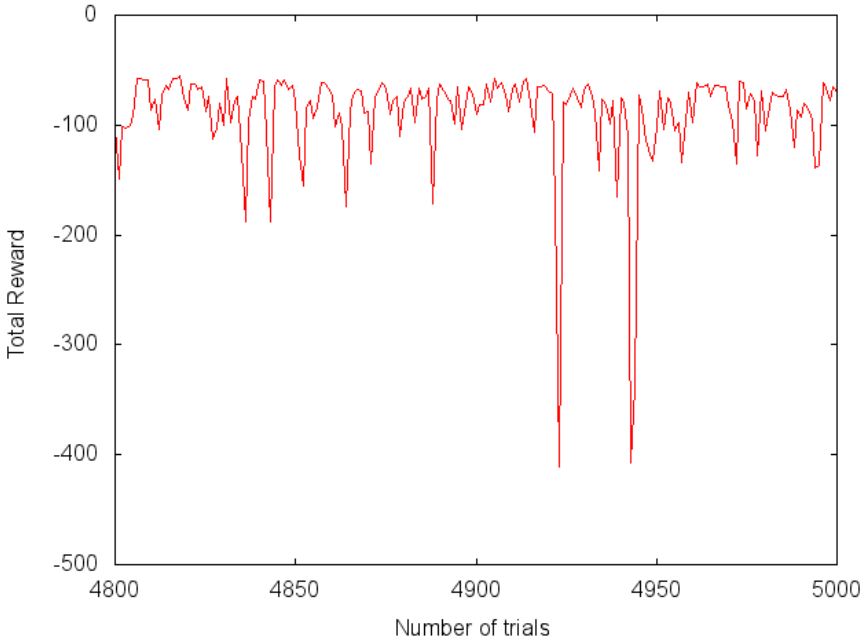
(a)：従来手法



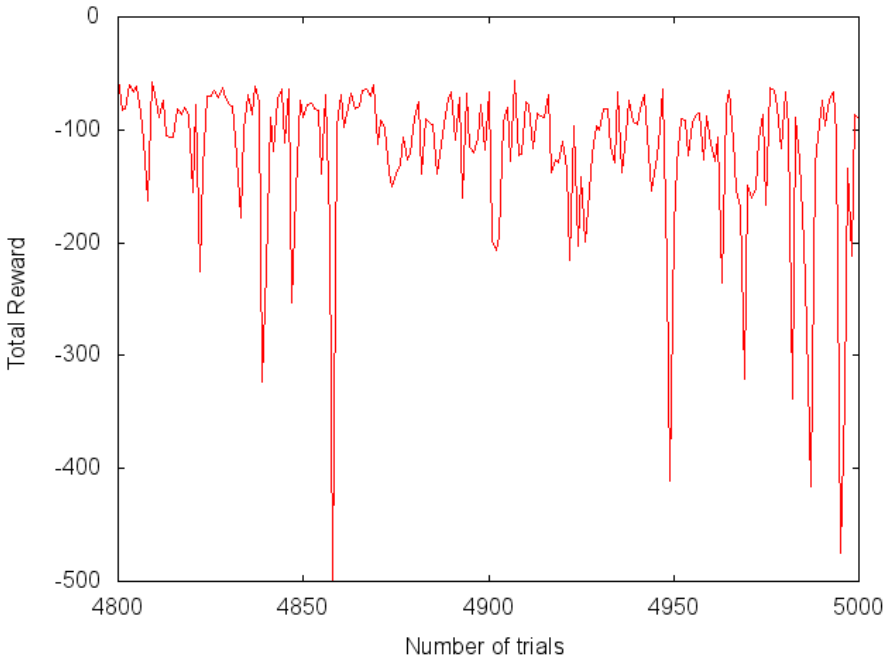
(b)：提案手法

図 20：図 19 の試行数が 0 回目から 200 回目の範囲の拡大図

次に図 19 の試行数が 4800 回目から 5000 回目の範囲を拡大したグラフを図 21 に示す．この試行数では従来手法，提案手法共に収束しているため，収束後の値の変化を比較する．図 21 を見ると，従来手法を用いたロボットと提案手法を用いたロボットの総獲得報酬の値は同じ値に収束していることが分かる．この結果から提案手法は従来手法と同じ総獲得報酬の値を得ることができることがわかる．一方で提案手法を用いたロボットは従来手法を用いたロボットに比べて各試行の総獲得報酬の値の差が大きくなっていることが分かる．この結果から提案手法を用いたロボットは従来手法を用いたロボットより収束後の各試行の総獲得報酬の値の差が大きくなることが示された．これは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しているためロボット全体がランダムに行動する確率が従来手法より高くなっているためである．



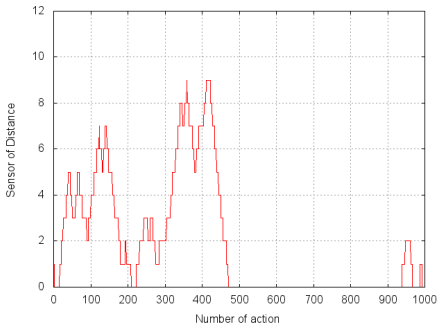
(a)：従来手法



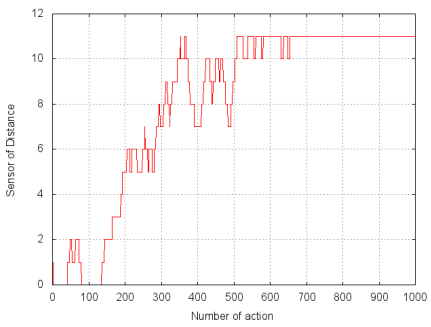
（b）：提案手法

図 21：図 19 の試行数が 4800 回目から 5000 回目の範囲の拡大図

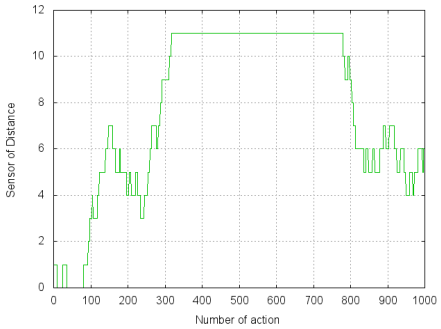
次に試行毎の台車のスタート地点からの走行距離の変位を示す．図 22 は台車ロボットの位置センサから取得した状態値の変位を，図 23 では台車ロボットの実際の走行距離の変位を示す．横軸は図 22，図 23 共に行動数．縦軸は図 22 では距離センサの値，図 23 ではスタート地点からの距離である．今回の実験では目標地点を $x_G = 5.5$ と設定してあるため状態値 x が 5 と 6 の間, 実際の距離では 50m と 60mの間を往復していれば収束している状態といえる．図 22 と図 23 を見ると，従来手法，提案手法共に 100 試行目までには目標地点に向かって収束していることが分かる．また 50 試行目を比較すると従来手法を用いたロボットは 300 から 400 行動の間で目標地点に収束しているが 提案手法を用いたロボットでは 100 行動前には目標地点に収束していることが分かる. この結果から提案手法を用いたロボットは従来手法を用いたロボットより少ない試行数で，目標地点に到達する行動を獲得していることが示された．これは台車のエージェントが目標地点に到達するための行動を出力できることを表す.



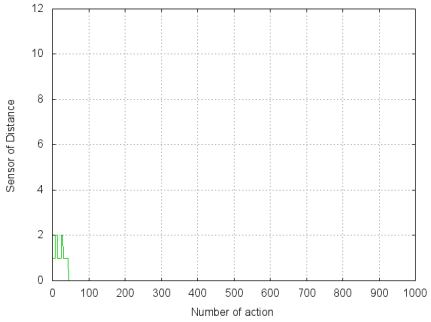
（a）：従来手法，1 試行目



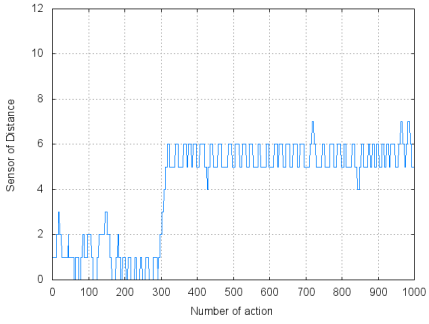
（b）：提案手法，1 試行目



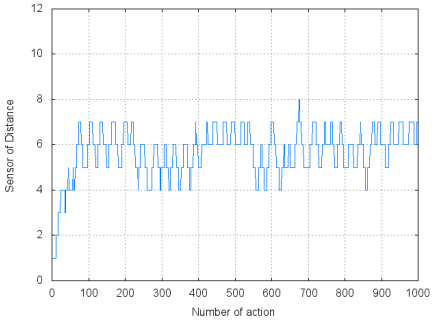
（c）：従来手法，10 試行目



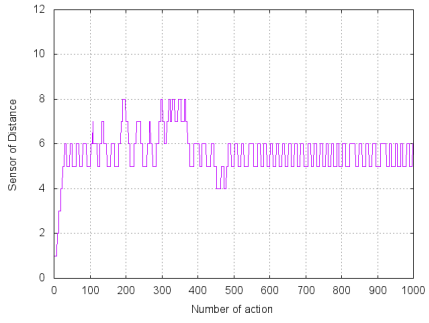
（d）：提案手法，10 試行目



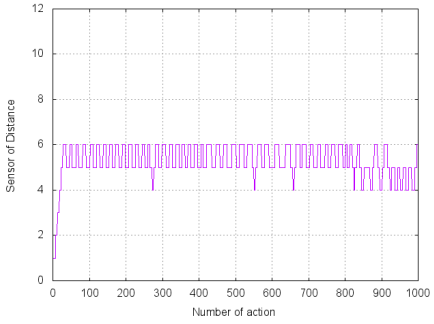
（e）：従来手法，50 試行目



（f）：提案手法，50 試行目

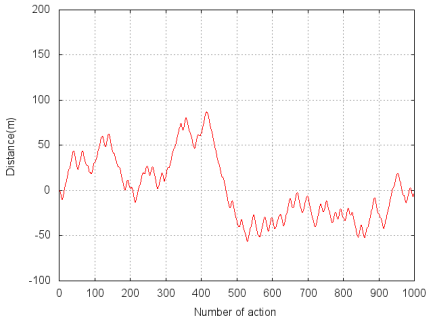


（g）：従来手法，100 試行目

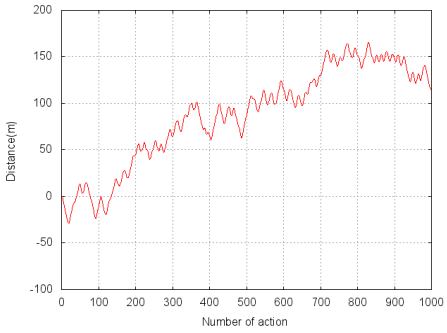


（h）：提案手法，100 試行目

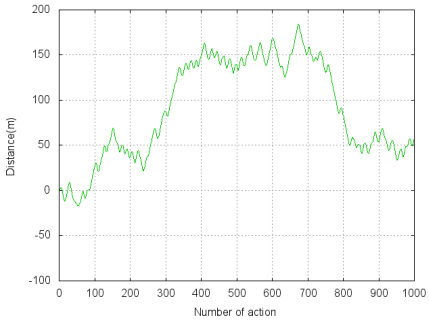
図 22：各手法の各試行の行動数と距離センサの状態値の変位



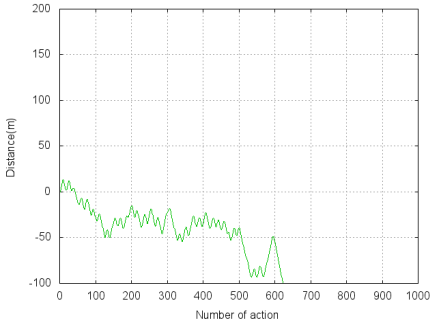
（a）：従来手法，1 試行目



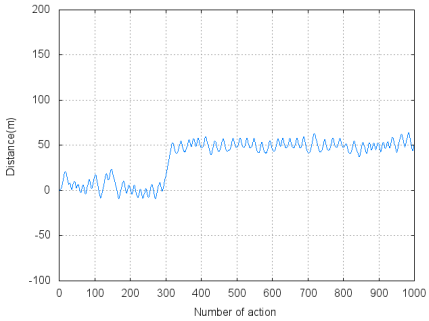
（b）：提案手法，1 試行目



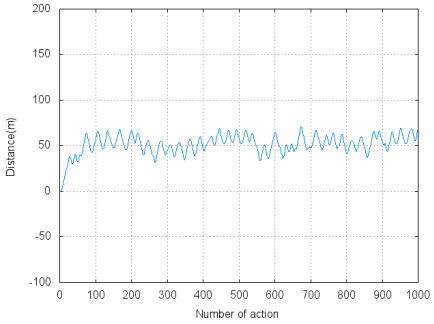
（c）：従来手法，10 試行目



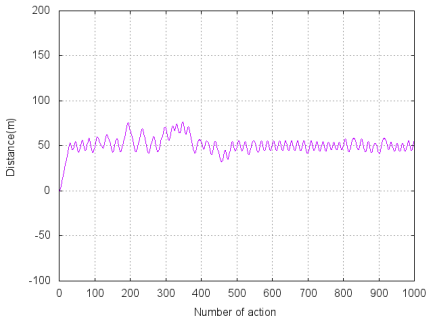
（d）：提案手法，10 試行目



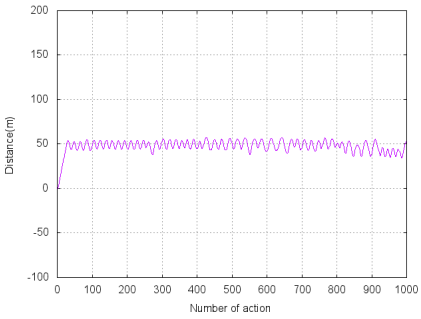
（e）：従来手法，50 試行目



（f）：提案手法，50 試行目



（g）：従来手法，100 試行目

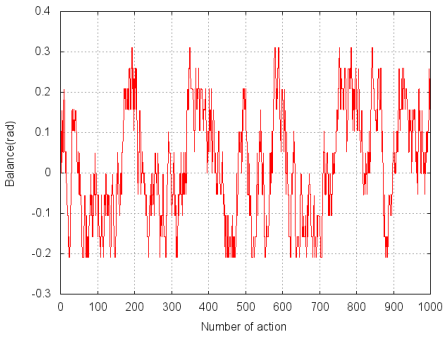


（h）：提案手法，100 試行目

図 23：各手法の各試行の行動数と走行距離の変位

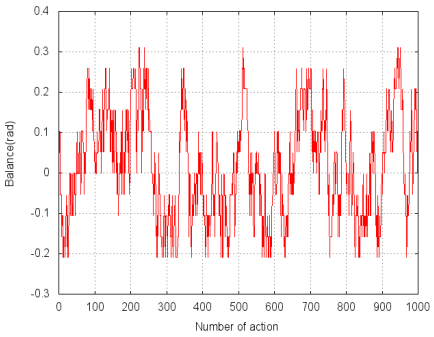
次に試行毎の台車ロボットに搭載されたテーブルの角度と台車の加速度と重力との合力のつりあいの変位を図２４で示す．横軸は行動数，縦軸は（テーブルの角度）－（合力の角度）でありラジアンで出力している．そのため 2 つの角度の差がないとき，つまり出力結果が 0 のときにテーブルの角度と合力がつりあっている状態にあるといえる．図 24 を見ると，従来手法，提案手法共に 100 試行目までにはテーブルの角度と合力の角度との差が 0 に収束していることが分かる．また 50 試行目を比較すると従来手法を用いたロボットは 300 から 400 行動の間で値が 0 に収束しているが，提案手法を用いたロボットは 100 試行目までには値が 0 に収束していることが分かる．この結果から提案手法を用いたロボットは従来手法を用いたロボットより少ない試行回数でテーブルの角度がものを落とさないための角度を保つ行動を獲得していることが分かる． これは台車のエージェントとテーブルのエージェントが協調動作を正確に行いお互いの状態に合わせた行動を学習していることを示す．

$\theta - R$

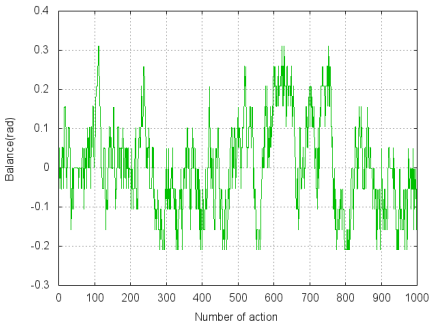


（a）：従来手法，1 試行目

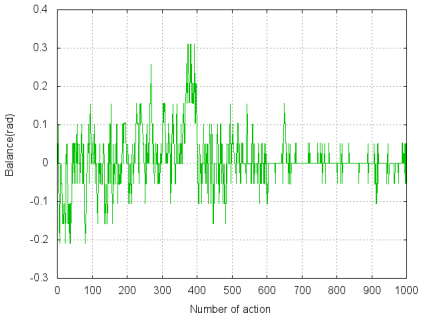
$\theta - R$



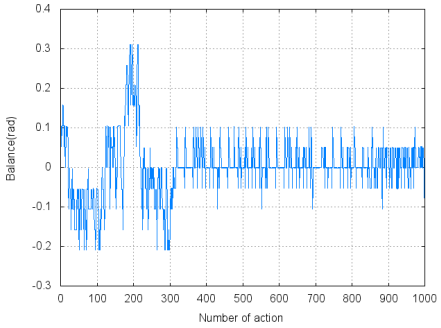
（b）：提案手法，1 試行目



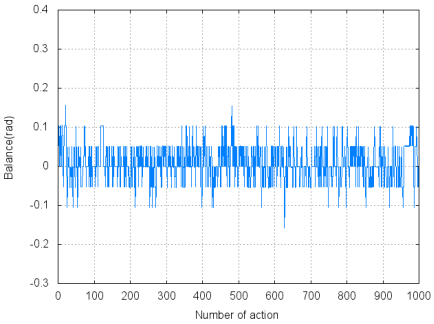
（c）：従来手法，10 試行目



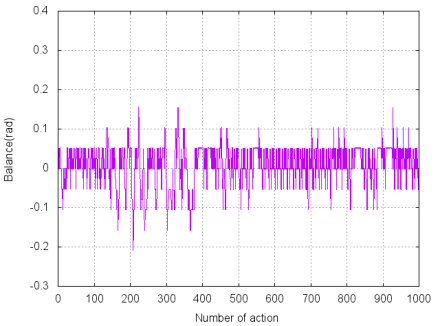
（d）：提案手法，10 試行目



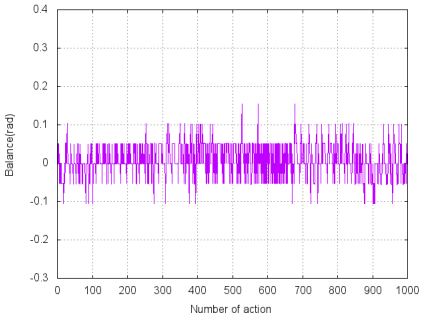
（e）：従来手法，50 試行目



（f）：提案手法，50 試行目



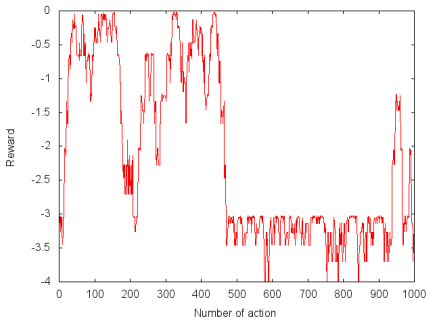
（g）：従来手法，100 試行目



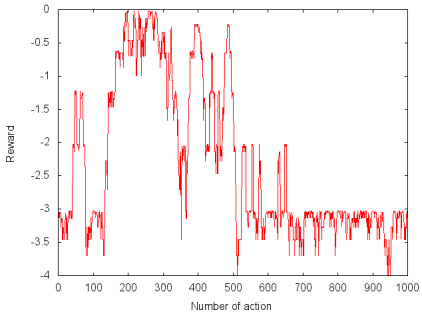
（h）：提案手法，100 試行目

図 24：各手法の各試行のテーブルの角度と合力のつりあいの変位

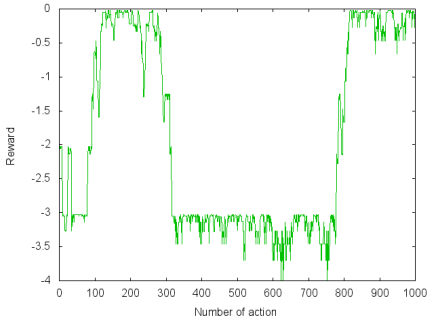
最後に各手法の各試行の行動数と獲得報酬の変位を図 25 に示す．横軸は行動数，縦軸は獲得報酬の値である．今回の実験での報酬の最高値は 0 と設定しているため，最終的に得られる報酬の値が 0 に収束していれば学習が収束している状態にあるといえる．図 25 を見ると従来手法，提案手法共に 100 試行目までには報酬が 0 に収束していることが分かる．また 50 試行目を比較すると従来手法では 300 から 400 行動の間で報酬が 0 に収束しているのに対して，提案手法では 100 行動前には報酬の値が 0 に収束していることが分かる．この結果から提案手法を用いたロボットは従来手法より用いたロボットより少ない試行数で獲得報酬の値が収束することが示された．一方で 100 試行目を比較すると従来手法を用いたロボットの方が提案手法を用いたロボットより，収束した後の行動毎の獲得報酬の値の差が少ない．この結果から提案手法を用いたロボットは従来手法を用いたロボットより収束後の獲得報酬の値が不安定になることが示された．



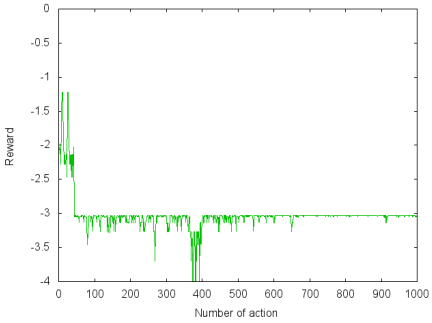
（a）：従来手法，1 試行目



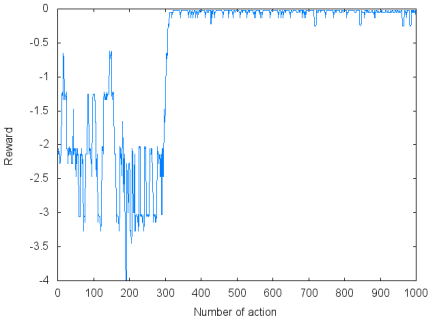
（b）：提案手法，1 試行目



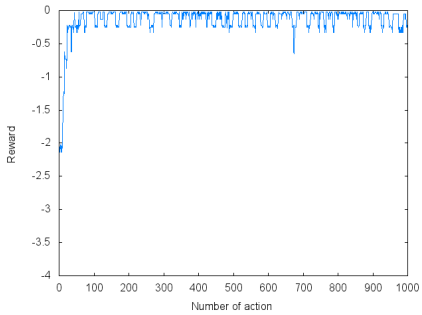
（c）：従来手法，10 試行目



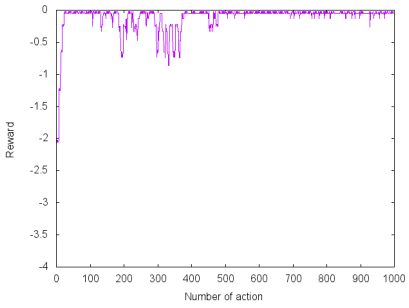
（d）：提案手法，10 試行目



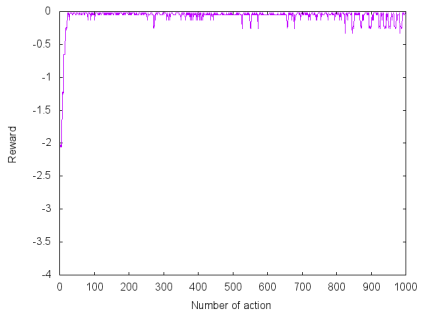
（e）：従来手法，50 試行目



（f）：提案手法，50 試行目



（g）：従来手法，100 試行目



（h）：提案手法，100 試行目

図 25：各手法の各試行の行動数と獲得報酬の変位

4.4 考察

各手法の 1 試行毎の総獲得報酬の推移のグラフから提案手法は学習が収束していることが分かる．また各手法の各試行の行動数と走行距離，テーブルの角度と合力のつりあい，獲得報酬それぞれの変位のグラフから提案手法は従来手法と同様の行動を獲得し適切な行動を得ていることが分かる．これは提案手法の各エージェントの協調動作が正しく作用し学習が収束しているためである．したがって提案手法は強化学習を正しく行えるシステムであることが示された．

またいずれの結果においても提案手法では従来手法より少ない試行数で学習が収束していることが分かる．これは提案手法の各エージェントの状態行動対が重複せずに分けられ，同時に学習を行えていることを示している．したがって提案手法は従来手法よりも学習時間を削減することができる手法であるといえる．しかし各結果から提案手法は従来手法より学習の精度が劣るという結果も同時に示された．原因としては，本実験では ε -greedy 法を使用しており，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するためであると考えられる．各エージェントが ε の確率でランダムに行動するためロボット全体がランダムに行動する確率は ε ×（エージェント数）となるため従来手法よりもランダムに行動する確率が高くなるためである．また各エージェントは他のエージェントの状態値は認識しているが出力する動作は認識していない．そのため各エージェントから見ると行動後のロボットの状態は自身の動作だけでは決められないため受け取る報酬に差が出る．この 2 つの問題が複合することで提案手法の学習精度の低下が発生していると考えられる．

この実験結果から提案手法は学習精度よりも学習速度の速さが望まれるタスクに適したシステムであるといえる．また学習精度がどれだけ必要となるかはタスクによって異なる．そのため若干学習精度が低下しても，タスク達成には影響しないタスクでは提案手法を用いるのが適しているといえる．

第 5 章　まとめ

5.1　論文全体の考察

本研究では，機械学習の問題点の 1 つである学習に多大な時間を要する点を解決するシステムの構築を目標とした．特に，強化学習において状態行動対の増加により学習時間が増大している点に注目した．状態行動対とは Q-learning における行動価値関数を構成する要素である状態と行動の組みのことである．そこで本研究ではマルチエージェントシステムを使用して状態行動対を複数のエージェントに分割して学習することで学習時間を短縮するアプローチを取った．強化学習をマルチエージェントシステムで行うにあたって，本研究ではロボットの行動に注目した．ロボットの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が協調動作することで生成される．そこで本研究では各アクチュエータにエージェントを設定し，アクチュエータ毎のマルチエージェント強化学習システムを提案した．

アクチュエータ毎のマルチエージェント構成で正しく学習を行えるために，他のエージェントの状態認識と，全アクチュエータの動作の同期を行った．他のエージェントの状態認識とは各エージェントが行動選択や学習を行う際に自身を含めた全アクチュエータの状態を取得することである．これにより各エージェントは現在のロボットの状態を正確に認識することができる．全アクチュエータの動作の同期とは各アクチュエータが実際に動作する際に，全アクチュエータで一斉に処理を行うことである．これにより各エージェントが取得する状態値が共通となる．

本研究では提案手法による学習が正しく行われていることと，目標である学習時間の短縮の達成の検証を目的にシミュレーション実験を行った．実験ではテーブルが搭載された台車ロボットの荷物運搬タスクによる比較実験を行った．台車ロボットには 2 つのアクチュエータが搭載されているため，提案手法では 2 つのアクチュエータにエージェントを設定した．タスクの特徴として台車の動作とテーブルの動作の協調動作によりタスクの達成度が変化する点がある．そのため提案手法では 2 つのエージェントが正しく協調動作しなければタスクを達成できないものとなっている．

実験の結果から，提案手法を用いたロボットは学習が収束し，従来手法を用いたロボットと同じ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は従来手法を用いたロボットより少ない試行回数で学習が収束した．このことから提案手法が強化学習を正しく行うことができ，学習時間の短縮を達成したことを示すことができた．しかし実験結果から，提案手法を用いたロボットは従来手法を用いたロボットより学習精度が低下することも示された．したがって提案手法は多少の学習精度の低下はタスク達成に対して問題とならず，学習速度の速さが望まれるタスクにおいて適しているシステムだといえる．

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

本研究では機械学習の中でも強化学習に的を絞り，その中でも Q-learning 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だと証明されているのは強化学習の中でも Q-learning だけであるのが現状である．しかし，アクチュエータ毎の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，更には他の機械学習手法にも適用できる可能性が存在する．提案手法が他の機械学習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．

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学習効果は保障されていない．また実ロボットに適用した際，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が発生する可能性がある．実ロボットによる実験と，実ロボットに適用した際の問題を発見し，それを解決することが必要となる．

5.2.3 学習精度の低下

本研究の実験で示したとおり，提案手法では従来手法より学習精度が低下する問題点が存在する．学習速度か学習精度のどちらを重要視するかは，タスクによって異なる．そのためタスク応じて使用する手法を使い分ければこの問題は発生しない．また学習精度が低下したとしてもタスク達成には影響しないタスクも存在する．そのため学習精度の低下は必ずしも深刻な問題とならない．しかし学習速度と学習精度の両方が望まれるタスクではこの問題が深刻となる．また問題とはならなくても学習精度は高いほうが望ましい．したがって学習精度の低下は解決なければならない問題の 1 つである．以下に学習精度の低下を引き起こす原因とそれぞれの解決方法の 1 つの案を提示しておく．

（a）　ε -greedy 法

本研究では強化学習の行動選択手法に ε -greedy 法を用いた．また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するように設定した．そのためロボット全体がランダムな行動を出力する確立は ε ×(エージェント数)となる．したがって提案手法は従来手法と比べてランダムに行動する確率が上がるため学習精度が低下する．

考えられる対策としては他の行動選択手法を適用する方法が挙げられる．またその他に

もエージェント毎にεの値を変える方法が考えられる．アクチュエータ毎に探索を優先するものと最適な行動を優先するものでεの値を変えれば探索とタスク達成のトレードオフのバランスの改善が期待できる．

(b)　各アクチュエータの動作連携

提案手法ではエージェントは各アクチュエータの状態を認識することでより適切な行動を獲得している．しかしエージェントが同じ状態で同じ行動を出力しても他のエージェントが出力する行動によって行動後のロボットの状態や受け取る報酬の値は異なる．そのためε-greedy 法の問題と複合することで最も高い報酬を得続けることを難しくしている．

考えられる対策としては各エージェントが他のエージェントの出力する動作を把握できることである．しかし各エージェントが同時に学習を行うため，他のエージェントが出力する動作を状態として認識することはできない．そこで過去の各エージェントが出力した動作の情報を利用する方法が考えられる．過去の経験情報を蓄積し他のエージェントが出力する可能性の高い動作を状態値として認識することで，獲得報酬を一定に保つことが期待できる．しかしこの方法では各エージェントが取得する状態値に新たな状態を加えることになるため，学習速度が悪くなる可能性がある．この考えを用いる場合は学習速度とのバランスが重要となる．

参考文献

[1] 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，No.1，pp51-56，1995

[2] 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，[知能と複雑系]，pp89-94，2004

[3] Watkins, C.J.C.H and Dayan, P.: Technical Note: “Q-Learning”, Machine Learning, Vol.8，pp.279-292，1992

[4] 山口智浩，増渕元臣，田中康祐，谷内田正彦，“経験型強化学習における仮想個体から実ロボットへの学習行動の伝播”，人工知能学会誌，Vol.12，No.4，pp570-581，1997

[5] S. P. Singh and R. S. Sutton：“Reinforcement Learning with Replacing Eligibility Traces”，Machine Learning，Vol.22，pp.123-158，1996

[6] R. S. Sutton：“Dyna, an Integrated Architecture for Learning, Planning, and Reacting”，Working Notes of the AAAI Spring Symposium, pp.151-155，1991.

[7] R. Maclin and J. W. Shavlik：“Creating Advice-Taking Reinforcement Learners”, Machine Learning，Vol.22，pp.251-281，1996.

[8] M. Tan：“Multi-Agent Reinforcement Learning：Independent vs. Cooperative Agents”, Proc. Of the 10th International Conf. on Machine Learning, pp.330-337, 1993.

[9] R. M. Kretchmar：“Parallel Reinforcement Learning”，The 6th World Conf. on Systemics，Cybernetics，and Informatics，2002.

[10] P. Cichosz and J. J. Mulawka：“Fast and Efficient Reinforcement Learning with Truncated Temporal Differences”，Proc. of the 14th International Conf. on Machine Learning，pp.99-107，1995.

[11] 片山晋, 小林重信, “TD(λ)学習の対数時間更新算法”, 人工知能学会誌, Vol.14, No.5, pp.879-890，1999.

[12] A. M. Printista, M. L. Errecalde and C. I. Montoya : “A Parallel Implementation of Q-Learning Based on Communication with Cache”, Journal of Computer Science and Technology, Vol.6, 2002.

[13] Mori, K. and Yamana, H. “A Fast Learning Method for Reinforcement Learning on Shared Memory Multiprocessors, Technical Report of IEICE(The Institute of Electronics. Information and Communication Engineers), Vol.2004, No.29, pp89-94, 2004”

[14] Mori, K. and Piat, E. “Parallel Learning Methods of Reinforcement Learning on Shared Memory Multiprocessors”, FIT2004, pp.291-292.

[15] Iima, H. and Kuroe, Y. “Swarm Reinforcement Learning Algorithm Based on Exchanging Information among Agents, Transactions of the Society of Instrument and Control Engineers”, Vol.42, No.11, pp1244-1251, 2006

[16] 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

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

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，中南義典さん，宮崎愛央さん，梅津祐介さん，北山直樹さん，渋谷和さん，杉本大志さん，沼田利伸君，三浦丈典君に感謝します．