平成　18 年度

卒 業 研 究 論 文

題　　目　痛覚を組み合わせた実ロボットの行動学習

提 出 者　　室蘭工業大学 情報工学科

氏　名　　尾上　由希子

学籍番号　　1523201

提出年月日　　平成　19 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　本研究の背景

現在，ロボットは工場などで稼働する産業用から娯楽や家事補助を目的とした家庭用まで，幅広く世の中に普及しつつある．特に近年家庭用ロボットは数多く開発／販売され，ロボットが我々人間にとってより身近な存在となりつつある．

多くの家庭――すなわちロボットのために整備された場所ではなく動作する環境が一様ではない所――にロボットが普及するにあたって，ロボットには様々な環境において使用者の期待する動きを実現しその性能を発揮する事が望まれる．例えば脚式ロボットの歩行動作一つをとっても，実験室から出た世界には路面の凹凸や坂道から雪道まで様々な地面が存在する．そのような場所で上手く歩き移動する事が脚式ロボットには望まれるであろう．

そういった一般的な生活環境において性能を発揮できるロボットの実現のために，あらゆる環境を人間が想定し，その環境にあった動作をロボットにプログラムするという手段がまず考えられる．しかし日常生活環境は時間と共に常に移り変わるものであり，多様で複雑なものである．したがって人間が全ての環境を予測しうる事は不可能であろうし，全ての環境を網羅する事には無理がある．

こういった，ロボットの置かれた環境への適応や人間がロボットをハードウェア／ソフトウェア両面において構築する際の負荷の軽減を考えると，ロボットが自ら学習し環境に応じた行動をとる事が望ましいと考えられる．また人間をはじめとする多くの生物が環境から得られる情報と経験（学習）によって行動を獲得している．そこでロボットにおいても環境から得られる情報とその学習によって行動を獲得するというのは望ましいものであるだろう．

1.2　従来研究

ロボットが環境に応じた行動を学習しその環境に適応していくためのアプローチとして，にコンピュータにおける機械学習・人工知能を用いた研究が挙げられる．

コンピュータに人間のような学習機能を持たせる事を目的とした機械学習には，大別して２種類の学習方法がある．学習のための正解情報が与えられている“教師あり学習”と正解情報が与えられず試行錯誤を通して学習を行う“教師なし学習”である．

前者の具体例としてニューラルネットワーク[3]を挙げる．ニューラルネットワークは入力データに対して理想的と考えられる出力値（正解情報）が与えられている際に，ニューラルネットワークからの出力と正解情報を比較することによって、その差をできるだけ小さくするよう結合荷重の値を変更し学習を行うというものである．このように教師

あり学習では，学習すべき正解情報と現状との誤差を拠り所として学習を進める．したがって人間から与えられる正解情報の無い世界，あるいは人間にも想像できないような環境においては適用する事が困難となる．日常生活環境は前述のとおり多様で複雑なものであるため，人間が正解情報を全て定義する事は困難である．それゆえ，ロボットが環境に適応するための手法として教師あり学習を適用する事は難しい．また，理想的な出力として与えられる正解情報は，学習するタスクに特化し依存した関数である．したがってタスクが変わると正解情報が変わりプログラムを作り変えなければならない事になる．一般的な生物が自然に行っている学習の方法と比較しても，正解情報が与えられそこから学習するというのは，何が正解であるか明瞭でないまま自発的な試行錯誤により学習を進めるという生物における学習とは異なっている．

後者の“教師なし学習”としては強化学習[1]が挙げられる．教師なし学習では，正解までの道筋・行動が明示される事はない．前述のニューラルネットワークのようにタスクに特化した正解情報が与えられる事も無く，学習者（エージェント）が環境との相互作用を行う事によりその時の行動の評価値のみが示される．環境との相互作用を繰り返す中で，その評価値を向上させるべく学習を行うという学習方法である．この，環境との相互作用による学習という手法自体が，環境に適応するためのロボットの行動学習と合致すると考えられるため，ロボットの学習において用いられる例が増えている[6][7]．

しかしこの学習法において，環境とのインタラクションから得られる評価値は主に人間によって設計される事が多い．この場合，学習する行動の評価は人間によって定義されたものであって，環境からの応答のみにより学習を行うという本来の意味での“教師なし学習”とは異なるものである．つまり，完全に環境からのフィードバックのみによって自分で自分の行動の評価を決定する機構が存在しない．

1.3　本研究の目的

人間など知的生物の学習の場合，他者から評価を与えられ行う学習と環境からのフィードバックによって評価を自分で決定する学習との２種類があると考えられる．前者は与えられた仕事の遂行など他者との関わり合いの中に生じるタスクについての学習であり，作業的なタスクの学習であると考えられる．一方後者は『転んだら痛いので坂道を転ばないように歩く』など自分自身の安全や活動維持ための学習であり，前者と比べてより原始的・本能的で，人間のみならず生物全般に共通する学習であるといえるだろう[4]．

これら２種類の学習は，人間が行動を行う際に双方ともに用いられているものであると考えられる．人間が他者に依頼された作業（タスク）を行う際，作業を完成するという目的と自分の活動を継続させる・安全確保という目的の２つを持ち行動を行う．タスク遂行の中で人間は作業の手順に関する学習・効率化と危険回避に関する学習・効率化

を同時に行っている．これは前者が依頼人から与えられた作業（タスク）及び目的に依存しそれに特化された学習であるのに対し，後者の学習は生命維持という人間にとっての本能であり永遠に継続される目的のためのもので，普遍的な学習である．前者のような目的依存性の強い学習と後者のような普遍的な学習というレベルの異なる学習・行動を同時に行うことによって，人間は自らの安全を確保しつつタスクの遂行をすることが可能となっていると考えられる．

これに対し，ロボットにおいて行われている学習は他者からの評価による学習しか無く，ロボットによる評価に基づいたロボット自身のための学習が存在しない．現在のロボットの多くは人間から与えられた目的を遂行するための学習すなわち目的に特化した学習を行っている．しかし，ロボットの行動設計時のプログラムの手間や環境の変化を考慮すると，人間のようにロボットが「動き続ける」という目標のために自律的に自らの破損を避けながらタスクを遂行する事が本来望ましいものである．その実現のため，人間の行う作業の学習と本能の学習の同時遂行のような，レベルの異なる学習をロボットに適用し即時的な目的に特化した学習と普遍的な目的のための学習を同時に行うシステムが求められると考えた．

そこで，本研究ではロボットにおいて即時的な目的に特化した学習と普遍的な目的のための学習を同時に行うシステムの構築を目標として，人間から与えられたタスクの遂行と自己の破損防止を同時に行うロボットシステムについて作成・実験し考察するものとする．

具体例として，人間にとって本能的評価の１つとなる「痛覚」をロボットについて定義・実装し，人から与えられたタスク学習と痛覚による危険回避学習というレベルの異なる学習を同時に行い実行するロボットの作成を目的とし実験を行う．

1.4　本論文の構成

以下に，本論文の構成を述べる．

第 1 章では，本研究の背景及び従来手法を述べ，本研究における目的を示した．

第２章では，本研究で用いる強化学習に関して，概念や手法及び各強化学習手法について説明する．また，コンピュータ上のシミュレーションで行った強化学習の適用事例を挙げる．

第３章では，本研究の実ロボットの実験において使用するロボットについて説明する．

第４章では，第 3 章で挙げたロボットに強化学習を適用し実際に学習を行わせた予備実験に関して，その内容と結果を述べる．

第５章では，本研究で提案するシステムのための「痛覚」を実装したロボットに関して，概念及び学習方法について説明する．

第６章では，本研究で提案する痛覚を組み込んだ学習について実ロボットにおいて行った実験の内容及び結果を述べる．

第７章では，本研究全体に関する考察及びまとめを行い，今後の課題を示す．

第２章　　強化学習について

本章では，本研究における学習手法として用いた強化学習[1]について説明する．

2.1　概要

強化学習は，学習者（エージェント）がエージェント外部の全てから構成される「環境」との相互作用を通して学習し目標を達成するという学習方法である．

強化学習手法が他の機械学習と大きく異なる点は，

(1) 学習に際して，正解が与えられない（教師なし学習）

(2) 学習する内容が，学習者の行動に依存する（能動性）

にある．したがって，

(1)どういう行動が望ましいかを予め明確化する必要が無く，

(2)ロボット自身が，学習すべき内容を能動的に決定し，

学習を進めることが可能となることが，最大の特長である．

強化学習では，環境内で観測・判断・行動するエージェントが，その行動の結果として受け取る報酬を最大化することを目標に学習を行う．

2.1.1　環境とエージェントの相互作用

強化学習におけるエージェント（学習者）と環境間の相互作用の概念図を図 2.1 に示す．環境から知覚される状態 S_t においてエージェントが何らかの行動 a_t を取った場合，環境からその行動の良し悪しを数値に写像した「報酬 r_t」を受け取り，他の状態に状態遷移を行う．その際得られた報酬を基にエージェントは学習を行う．

2.1.2　強化学習の構成要素

方策（行動選択手法）・報酬関数・価値関数（行動価値の評価・推定）・環境のモデル，これらが強化学習の主な構成要素である．これら構成要素及び強化学習で頻繁に用いられる単語について本節では説明を行う．

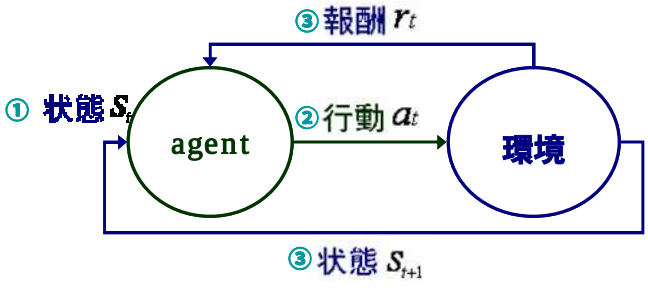


図 2.1　エージェントと環境の相互作用

・エージェント

強化学習の枠組みにおいて，学習と意思決定を行う者を示す．
ロボットの行動学習などの場合においてはロボットを指す．

・環境

エージェント外部の全てから構成され，エージェントが相互作用を行う対象である．

・方策（行動の選択方法）

ある時点での学習エージェントの振舞い方を定義する．方策は、環境において知覚した状態から、その状態にあるときに取るべき行動への写像である．この方策は一般的には確率的である．

・報酬関数

報酬関数は目的を定義する。エージェントがとった行動に対する評価を数値化したものであり，報酬はその状態におけるエージェントの行動の望ましさを表している．強化学習エージェントの唯一の目的は最終的に受け取る報酬を最大化することである．エージェントが報酬関数を変更することはできないが方策を変更する指針として使うことができる．報酬関数も一般的には確率的である．

・価値関数

報酬関数が即時的な行動の良さを表すのに対し、価値関数は最終的な行動の良さを指定する．状態の価値とは、エージェントがその状態を基点として将来にわたって蓄積することを期待する報酬の総量であり、その後に続きそうな状態群とそれらの状態群で得られそうな報酬を考慮に入れた上での長期的な望ましさを示す．

意思決定を行い、決定の結果を評価するには、価値に最も関心を払う．行動の選択は価値を判断した結果に基づいている．最大の報酬を得るためには報酬ではなく、最も高い価値を持つ状態につながるような行動を見つけ出そうとする行動が必要となる．故に強化学習問題では価値を評価・推定することが最も重要とされている．

2.1.3　強化学習の流れ

強化学習は，2.1.1 で示した相互作用により行われる．その流れを具体的に述べる．

強化学習の流れを図 2.2 に示す．環境から知覚した状態 S によって，エージェントは自分の行える行動の中から，その状態における行動価値に基づき行動選択手法を用いて行動を選択し実行する．その結果得られた報酬を基に，エージェントはその状態において選択した行動を行うことの価値（行動価値）の更新を行動価値関数によって行い学習し，次回同様状態においての行動選択に活かす．

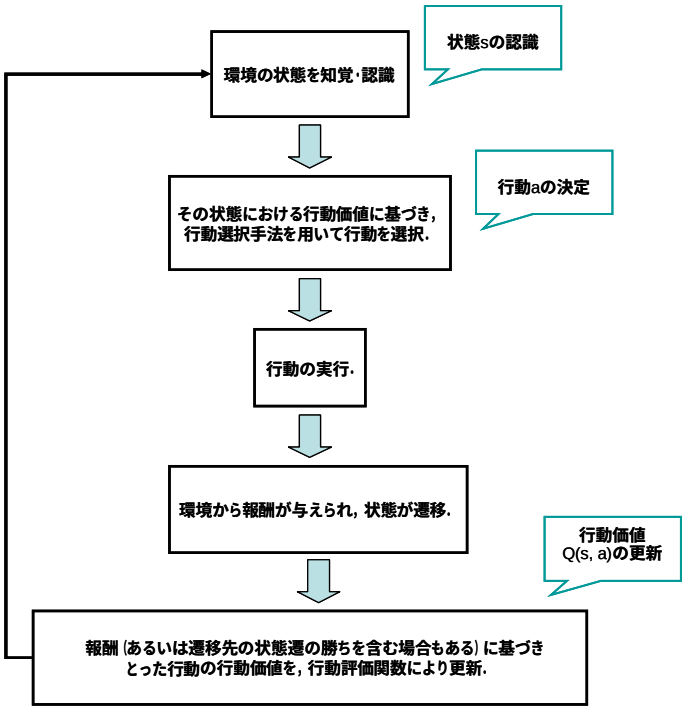


図 2.2　強化学習の流れ

2.2　行動選択手法と行動価値の評価方法

2.2.1　行動選択手法

行動選択手法とは，エージェントが認識した環境の状態 S においてとる行動を選択する際に，用いられる手法である．

強化学習における行動選択の際に重要となるのは，単に現在の推定価値（状態価値または状態行動価値）が最大となる行動を選択するのみでなく，より価値の高い行動を求める探索を行うことである．両者間のトレードオフを exploration-exploitation 問題という．探査を継続することは，局所的最適解に陥らずに方策の正しい価値推定を行うため，また非定常問題において環境の変化に追随するために有効である．

探索と知識（現在までに学習した内容）利用の両立という観点から，比較的よく用いられる行動選択手法として，ε-greedy と softmax 手法がある．以下にその行動選択手法を説明する．

ε-greedy 手法においては，推定される行動価値が最も高い行動（グリーディな行動）を(1-ε)の確率で選択する（exploitation に相当する）か，小さい確率εで一様に任意の行動を選択する（exploration に相当する）という手法である． εが小さいほど，最適な行動が行われる回数は多くなるが，最適な行動を見つけ出すまでに時間がかかってしまう．それゆえ，探索と知識利用のバランスの取り方を考える必要がある． ε-greedy 手法の欠点として，確率εにおける行動選択の際にほとんど最悪と思われる行動を選択する可能性とほと

んど最適行動に近い行動を選択する可能性が同じくらいに高くなるという事がある．

softmax 手法は，推定される行動価値に基づいた確率で行動を選択するという行動選択手法である．一般的に Gibbs 分布，あるいは Boltzmann 分布に基づいて行動が選択される．具体的には，t 回目の試行における行動 a の行動価値 $Q_t(a)$ が与えられた場合，行動 a を選択する確率 $\pi(a)$ は次式で与えられる．

$$\pi(a) = \frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}} \tag{2.1}$$

ここで τ は温度と呼ばれる正定数で，温度が高い場合には，すべての行動がほぼ同程度に起こるように設定され，低い場合には、価値の推定が異なる動作の選択確率の差がより大きく異なるように設定される．

2.2.2 行動価値の評価・推定手法

強化学習における行動価値の推定手法について説明する．

行動 a をとった際の平均報酬を行動 a の真の価値 $Q^*(a)$ とし，t 回目の試行におけるその推定量を $Q_t(a)$ とする．強化学習において，学習エージェントは行動 a の真の価値そのものを知ることはできず，行動によって得られる報酬からそれを推測した $Q_t(a)$ を学習し行動選択に用いる．この行動価値推定の方法の１つとして，標本平均化手法がある．

標本平均化手法は，その行動が選ばれたときに実際に受け取られた報酬を平均化してゆく方法である．t 回目の試行において，それまでの間に行動 a が k_a 回選択されていて，各回で得られた報酬が $r_1, r_{2,} \cdots, r_{k_a}$ とすれば，行動 a の推定価値 $Q_t(a)$ は次式で求められる．

$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a} \tag{2.2}$$

$k_a = 0$ の場合には，$Q_t(a)$ を，$Q_0(a) = 0$ のようなデフォルト値に設定する．$k_a \to \infty$ の極限において，大数の法則から，$Q_t(a)$ は $Q^*(a)$ に収束する．

本節で述べた行動選択手法と行動価値の推定手法の拡張が，様々な強化学習の手法となる．その強化学習手法の一般的な手法のうち，本研究で用いたものについて次節で述べる．

2.3　一般的な強化学習手法

本節では強化学習法においてよく用いられ，本研究においても使用した学習手法について説明する．

2.3.1　強化比較法

強化比較法は，状態遷移の無い比較的単純な強化学習課題に用いられる手法である．強化比較法では，与えられた報酬の大小を評価するための基準レベルをリファレンス報酬と呼び，現在までに受け取った報酬の平均値を用いる事が多い．

この基準レベルより大きい報酬が得られた行動は，良い行動と判断され以後この行動をとる確率が上がる．一方，基準レベルを下回る報酬につながった行動に関しては，以後この行動をとる確率を下げることにより次第に報酬の大きな行動が選択される傾向が強まる．

実際の行動選択に当たっては，通常 softmax 手法(2.2.1 参照)が用いられる．この場合，t 回目の試行において行動 a を選択する優先度 $p_t(a)$ を用い，行動 a を選択する確率 $\pi(a)$ は次式で与えられる．

$$\pi_t(a) = \frac{e^{p_t(a)}}{\sum_{b=1}^{n} e^{p_t(b)}} \tag{2.3}$$

また，行動 a を選択する優先度及びリファレンス報酬 $\bar{r}$ は次式によって更新される．

$$\begin{aligned} p_{t+1}(a_t) &= p_t(a_t) + \beta[r_t - \bar{r}_t] \\ \bar{r}_{t+1} &= \bar{r}_t + \alpha[r_t - \bar{r}_t] \end{aligned} \tag{2.4}$$

ここで，β は正のステップサイズ・パラメータを示し，優先度に関わる報酬の重みを表す．また α（0< α <1）はリファレンス報酬の学習率を示している．

2.3.2　追跡手法

追跡手法は，行動価値推定と行動優先度の両方を利用した学習手法である．優先度は現在の行動推定価値に従ったグリーディな行動を「追いかける」目的で使用される．t 回目の試行で行動 a を選択する確率 $\pi(a)$ を行動優先度として用いられる事が多い．

毎回の試行の直後，グリーディな行動が選ばれる可能性がより高くなるように，この確立値は更新される．t 回目の試行の後，t+1 回目の試行に対するグリーディな行動（複数個ある場合にはその中からランダムに選んだ１つ）を $a^*_{t+1} = \arg\max_a Q_{t+1}(a)$ とする．この場合，

行動 $a_{t+1} = a^*_{t+1}$ の選択確率は

$$\pi_{t+1}(a^*_{t+1}) = \pi_t(a^*_{t+1}) + \beta[1 - \pi_t(a^*_{t+1})] \tag{2.5}$$

で表され，確率 1 に向かってβの比率で増加させられる．残りの行動の選択確率は，全ての $a \neq a^*_{t+1}$ に対して，次のように 0 に向かって減少される．

$$\pi_{t+1}(a) = \pi_t(a) + \beta[0 - \pi_t(a)] \tag{2.6}$$

行動価値 $Q_{t+1}(a)$ は，標本平均化手法（2.2.2 参照）などを用いて更新される．

2.3.3　Q 学習

多くの強化学習手法は，離散化された状態空間と時間の上に組み立てられている．本論文で主に用いる Q 学習という学習法は，継続する状態間の効用の差分を利用することから，時間的差分学習（TD 学習）と呼ばれる強化学習手法に分類される．

時間的差分学習とは，環境のダイナミクスのモデルを用いずに経験から直接学習することができ，最終結果を待たずに他の推定値の学習結果を一部利用し，推定値を更新する学習法である．

Q 学習は，行動価値（ある状態である行動をとることの価値で，一般的に Q 値と呼ばれる）用いた，方策オフ型の TD 学習手法であり，ある方策（挙動方策と呼ばれる）に基づいて行動しながら，最適方策を学習する点に特徴がある．例えば行動選択手法として ε-greedy 手法を用いた場合，ε-greedy 手法に基づく行動決定を行いながら実際には最適方策を学習する．

1 ステップ Q 学習における行動価値の推定の改善は次式によって行われる．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{2.7}$$

ここで，s_t は現在の状態，a_t は採用した行動，r_{t+1} は行動によって得られた報酬を示し，s_{t+1} は行動後の新しい状態，a は新しい状態において選択される行動である．また，$Q(s_t, a_t)$ は状態 s_t における行動 a_t の行動価値推定を示し，α(0< α <1)は学習率，γ (0≦ γ <1)は割引率を表す．

2.4　n 本腕バンディット問題

この節では，強化学習の行動からの学習の例としてよく用いられる n 本腕のバンディット問題について述べ，シミュレーションを行う．

2.4.1　n 本腕バンディットとは

n 本腕バンディットとは，１台のスロットマシンにｎ本の腕（レバー）があり，腕ごとに当たりの出る確率が設定されているタスクである（図 2.3）．m 回の試行のうち，『当たり』の出る確率の１番高いレバーを発見できると高得点につながる．

１回の行動選択がスロットマシンの１つのレバーを引くプレイに相当し，報酬は当たりで得られる利益に相当する．スロットマシンのプレイヤーはプレイを繰り返しながら，最良のレバーを探し賞金を最大にするよう努力する．これはエージェントが報酬の最大化を目指し学習を行う事に相当する．

n 本腕バンディットタスクは，『状態が１つで選択できる行動がｎ個ある場合に，最善の行動を学習するタスク』と読み替えることができる．

2.4.2　n 本腕バンディットへの強化学習の適用

強化学習における『行動の学習』の具体例として，n 本腕バンディットタスクを用いシミュレーションを行った．

エージェントが人間と同様に，『当たり』を出す確率の高いレバーがどれであるかを何度かの試行の中から発見し，高得点が得られるようになるまで学習することを目的とする．

各試行においてエージェントはレバーの選択を行う．バンディットマシンから出た『当たり』『はずれ』を点数化し，報酬として与える．その報酬に基づいてどのレバーを選択すると高得点につながるかの学習をエージェントに行わせる．

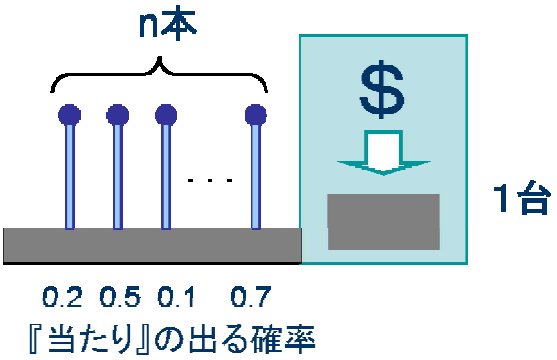


図 2.3　n 本腕バンディットマシン

これらを図 2.4 に示すアルゴリズムに基づき，ε-greedy 法及び softmax 法をそれぞれ行動選択手法として用い，行動価値推定に標本平均化手法を用いた学習方法・行動選択手法に softmax 法を用いた強化比較・行動価値推定に標本平均化手法を用いた追跡手法の４種類の学習手法を用いてエージェントに学習を行わせた．以下に，各学習手法をどのように用いたかを示す．

ε-greedy 法及び softmax 法をそれぞれ行動選択手法で用い行動価値推定に標本平均化手法を用いた学習方法の場合，バンディットマシンのレバーを選択する際にどのレバーを選ぶかという「選択方法」において行動価値を基に各行動選択手法を使用した．また，行動後得られた報酬を基に行動価値の更新を行う．この計算に，標本平均化手法を用いた．

強化比較を用いる場合は，バンディットマシンのレバー選択の際に強化比較法の特徴である「優先度」を基に softmax 法により選択するレバーを決定した．また行動後得られた報酬を基に，行動の優先度及びリファレンス報酬の更新を行った．

また，追跡手法の場合，バンディットマシンのレバー選択の際に追跡手法の特徴である「選択確率」に基づく確率的選択によってレバーを決定した．そして行動後得られた報酬を基に，標本平均化手法を用いた行動価値更新と選択確率の更新を行った．

2.4.3　シミュレーション設定

n 本腕バンディットマシンの腕の数 n を３本と設定し，それぞれに ARM0, ARM1, ARM2 と名付ける．また，各腕について，ランダムな確率で『当たり』を出すものとする．今回のシミュレーションにおける各腕の『当たり』を出す確率は以下のように設定した

．

ARM 0：　0.187454
ARM 1：　0.662203
ARM 2：　0.898992

```
行動価値など各種初期化.
入力された学習回数分繰り返し:
        各種選択手法による腕(レバー)の選択.
        N本腕バンディットマシンでplay(報酬を得る).
        得た報酬rを基に行動価値等各種更新.
学習回数に達したら終了.
```

図 2.4　n 本腕バンディットにおける学習アルゴリズム

また今回は『当たり』＝1点／『はずれ』＝0点と点数化する．その得点を 1 回の試行で得られる報酬として報酬を与え各試行において学習させる．

各学習手法を用いる際のパラメータは以下のように設定した．

ε-greedy 法：　ε=0.01

SOFTMAX 法：　τ=0.1

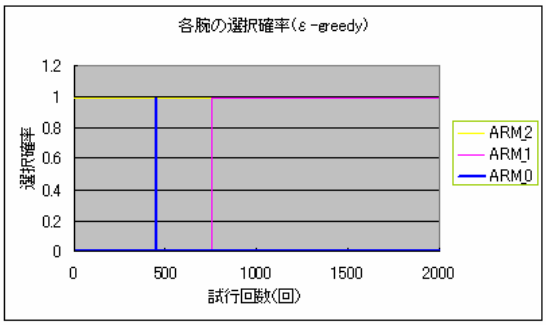
強化比較：　α=0.1，β=0.05

追跡手法：　β=0.05

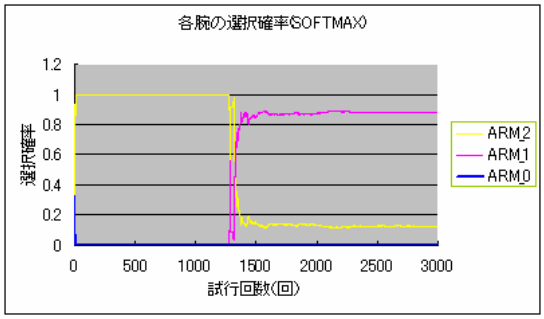
以上の設定のもと，それぞれの学習手法で各 2000 回試行を行った．

2.4.4　結果

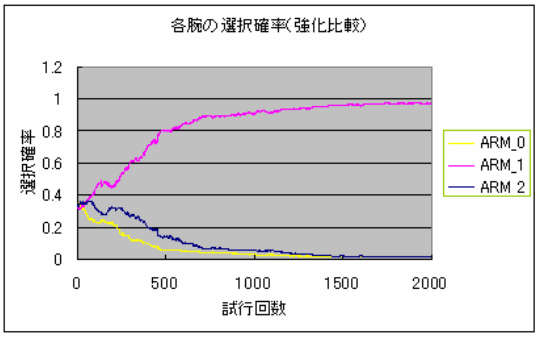
2.4.3 で示した各学習手法での試行における腕の選択確率の推移を以下の図 2.5 に示す．



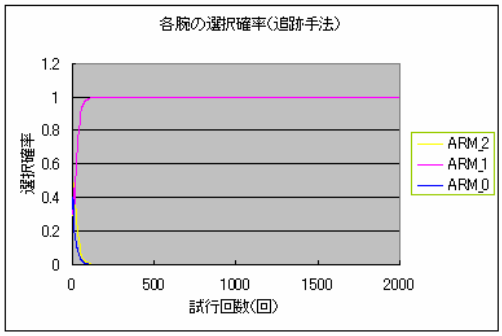
(a) ε-greedy 法を用いた場合



(b)SOFTMAX 法を用いた場合



(c)　強化比較を用いた場合



(d)　追跡手法を用いた場合

図 2.5　それぞれの手法を用いた場合のレバー選択確率の推移．

2.4.5 考察

図 2.5 に示した行動選択確率の推移のグラフより，どの手法を用いても最終的には最も『当たり』の出る確率の高い２番のレバーの選択確率が高くなるよう収束していることがわかる．

しかし，用いる手法によって，選択確率の推移の模様は異なる．ε-greedy 法の場合，その時に最適と思われる行動をメインとして選択を行うため今回の実験のように『当たり』を出す確率が最良に近いものが２番目にあると，しばらくの間そちらをメインに選択を行うことが分かる．softmax 法の場合もバンディットマシンが『当たり』を出すタイミングによってはしばらくの間２番目に『当たり』を出すレバーの選択確率が高くなっている事が分かる．さらに，他の手法と比べて最良のレバーの選択確率が高く収束した後も，2 番目に良いレバーの選択確率がある程度高くなっている．これは行動価値を基に選択確率を決定しているためであると考える．強化比較はかなりゆるやかに選択確率が収束していっているが，これは優先度の更新に用いるパラメータα・βを小さく設定したためであると考えられる．追跡手法は，各手法の中で最も早く選択確率が収束している．

2.4.6 まとめ

本節では，強化学習法によりエージェントが n 本腕バンディットマシンにおいて『当たり』の出る確率が高いレバーがどれであるかを学習するシミュレーションを行った．

行動選択手法及び学習手法としてε-greedy 法・SOFTMAX 法・強化比較・追跡手法を用いて試行を行い，それぞれにおいて学習が行われる事の確認及び各手法の比較を行った．

その結果，2.4.4 のシミュレーション結果より学習が行われたことを確認した．また，各手法において特徴があり，収束までの過程及び時間に差が出た．その結果，今回のシミュレーションにおいては追跡手法が最も早く収束する結果となった．

2.5　迷路問題

本節では，迷路問題のシミュレーションを通して強化学習の『状態と行動の学習』についての例を示す．

2.5.1　迷路問題とは

迷路問題とは，図 2.6(a)に示すような障害物がランダムに配置された 2 次元正方格子状のグラフからなる迷路において図 2.6(b)に示す行動が与えられている際に，初期状態（スタート地点）から目標状態（ゴール）への経路を求める問題である．

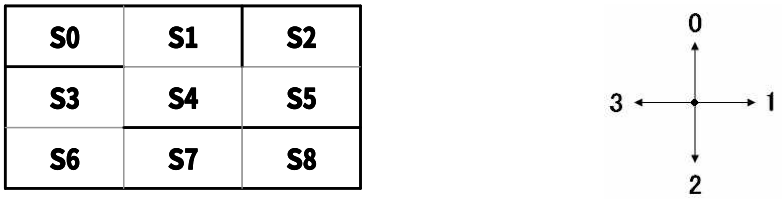
2.5.2　迷路問題への強化学習の適用

今回のシミュレーションでは，図 2.6(a)に示す迷路を，スタート：S0，ゴール：S8 と設定しスタートからゴールまでの道のりをエージェントに学習させることを目標とし，シミュレーションを行った．

各格子の中を一つの状態として，図 2.6 に示すように $S0 \sim S8 \in S$：状態，$A0 \sim A3 \in A$：行動と定義した．

報酬として，ゴールに到る状態遷移となる行動に対し正の報酬，それ以外の行動には負の報酬を与え，状態遷移毎に学習を行わせた．

学習手法には Q 学習法を用い，以下の図 2.7 で示すアルゴリズムで迷路問題の解決を行った．



(a)迷路（環境）　　(b)移動方向（行動）

図 2.6 迷路問題での環境／行動

```
すべてのs∈S, a∈Aについて Q(s, a)を初期化．
入力された学習回数分繰り返し：
        s←S0(スタート地点)
        s==S8 (ゴール) まで繰り返し：
                行動選択手法により，状態sでの行動aを選択．
                行動aの後，報酬rと次状態s'を観測．
                Q(s, a)の更新．
                s←s'
        sがゴールであれば繰り返し終了．
学習回数が入力された回数を満たせば繰り返し終了．
```

図 2.7　Q 学習法を迷路問題に適用したアルゴリズム．

2.5.3　シミュレーション設定

図 2.6 (a)に示すように，現在位置を状態として観測可能な迷路を設定する．

各状態での意思決定において，図 2.6(b)に示すような４種類の方向へ移動する行動のうち１つを選択し実行すると，その行動に応じた状態遷移を行う．状態遷移は選択した行動の示す方向の状態への移動となるが，障害物等により次の状態へ進めない場合はその状態に留まるという形になる．

ゴールに到達すると正の報酬（＋10）を与え，それ以外の状態の状態遷移では報酬は 0 を与えるよう報酬を設計する．

学習手法としてQ学習法，行動選択手法にε-greedy 法・softmax 法，また追跡手法の行動優先度の考え方を用いそれぞれシミュレーションを行う．

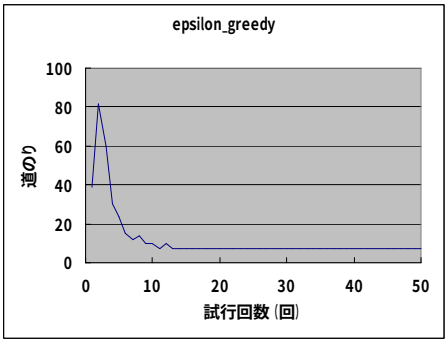
各パラメータの設定を以下に示す．

学習率		$\alpha = 0.1$
次状態の行動価値の重み		$\gamma = 0.5$
各行動選択手法において		
	ε-greedy 法	$\varepsilon = 0.01$
	softmax 法	$\tau = 0.1$
	追跡手法	$\beta = 0.05$

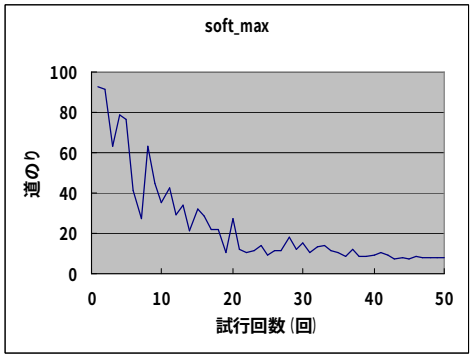
以上のように設定したパラメータを用い，それぞれの手法について 5000 回スタートからゴールに到るまでの試行を実行し，それを５セット行いデータの平均を取った．各試行回数においてゴールまでにかかる道のり（状態の移動の回数）と，全状態での行動価値の増分を取りそれらの平均によって学習が行われているかを検証した．

2.5.4　結果

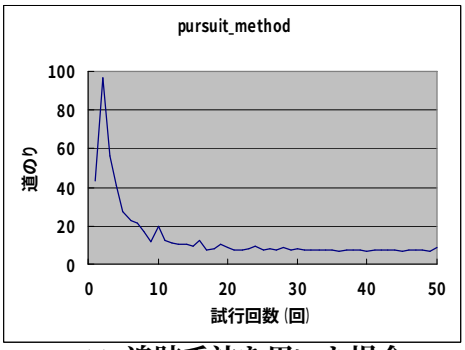
ε-greedy 法・softmax 法・追跡手法と組み合わせた Q 学習における，試行回数と道のりの平均のグラフ及び行動価値の増分のグラフを図 2.8・図 2.9 に示す．それぞれ収束までの特徴が出た部分として，試行回数と道のりの平均のグラフでは 50 回まで，行動価値の増分のグラフでは 500 回までの試行データをグラフ化した. それぞれ 50 回及び 500 回までの試行でグラフは収束し，それ以降値の変動は見られなかった．



(a) ε-greedy 法を用いた場合

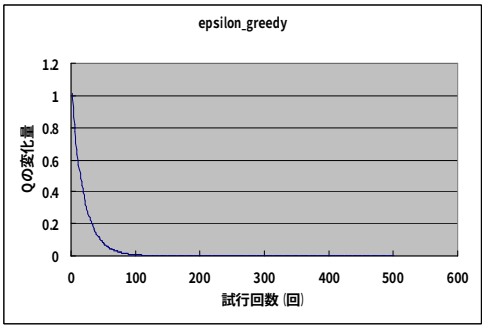


(b) softmax 法を用いた場合

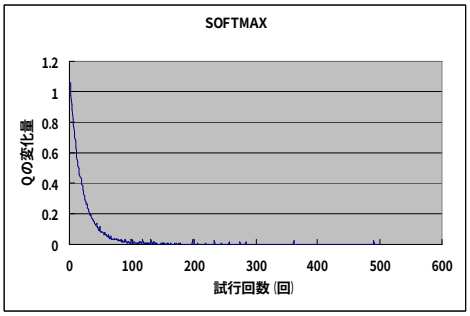


(c) 追跡手法を用いた場合

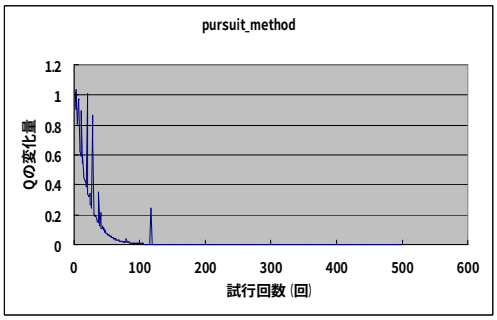
図 2.8　試行回数とゴールまでに要した道のりのグラフ.



(a) ε-greedy 法を用いた場合



(b)softmax 法を用いた場合



(c)　追跡手法を用いた場合

図 2.9　試行回数と行動価値の変化量のグラフ.

2.5.5　考察

試行回数とゴールまでに要した道のりのグラフより，どの行動選択手法においても道のりの数が最終的にゴールまでに要する最短の道のりの数である 7 に収束している事がわかる．これより，どの行動選択手法でもQ学習法による学習は行われることがわかる．

また行動価値の増分のグラフが最終的に 0 に収束している事より，推測されていた行動価値が一定の行動価値へと収束した事がわかる．

これらより，迷路問題においてエージェントは道のりを学習したと言える．

それぞれの行動選択手法について見てみると，図 2.8 より ε-greedy 法が最も早く行動が収束している事がわかる．softmax 法はなかなか収束していないが，これは softmax 法が行動価値を基に確率的に行動を選択するため行動価値に大きな差が無い場合は確率にもさほど差が出ないため最適以外の行動も選択される確率が高くなるからである．追跡手法は ε-greedy 法の結果とほとんど変わらないが， ε-greedy 法よりも収束が遅い．

2.5.6　まとめ

本節では，強化学習によりエージェントが迷路においてゴールまでの道のりを学習するシミュレーションを行った．

学習方法としてQ学習法を用いた．行動選択手法として ε-greedy 法・softmax 法・追跡手法という３種類の行動選択手法でそれぞれ試行を行い，学習が行われる事の確認及び行動選択手法の比較を行った．

その結果，2.5.4 のシミュレーション結果より学習が行われたことを確認した．また各行動選択手法をQ学習法で使用した際の特徴を示した．

2.6　まとめ

本章では，本研究において用いられる学習法である強化学習ついて説明し，その概要と用語及び各学習手法について述べた．

また n 本腕バンディット問題及び迷路問題を用い，強化学習のシミュレーションを行った．これらシミュレーションを行う上で，各行動についての評価値（報酬）は全て人間が設定した．状態遷移の無い，単発の行動学習である n 本腕バンディット問題においては『当たり』＝1 点／『はずれ』＝0 点と設定するのみであったが，迷路問題においては各状態遷移において報酬を設定する必要があった．それによって，特に状態遷移のある学習において，報酬を逐次設定する事に手間がかかる事がわかった．また，各シミュレーションにおいて，この章で示した行動学習法によって行動の学習が行われている事が確認された．

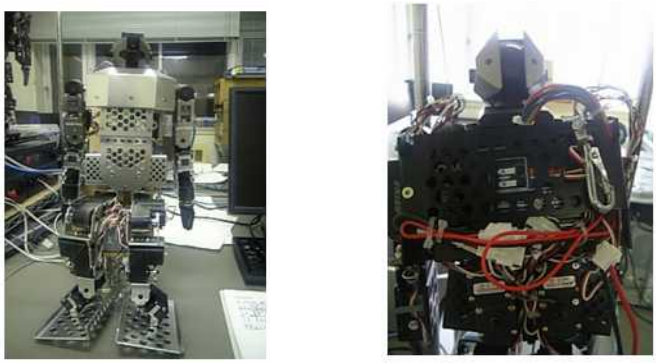
第 3 章　speecys ロボット

本研究における実ロボットの実験で用いたのは，speecys 社の提供する SPC-001 というロボットである．以下，このロボットについて説明を行う．

3.1 実験に使用したロボット：SPC-001

SPC-001 とは，speecys 社の提供する人型ロボットである．（図 3.1）

全身に21個のサーボがあり，各サーボにおいて負荷などの検出が可能である．また，背面にA/Bボタンがついており，押したか押されていないかを検出する事ができる．また，NetBSDをもとにロボット用にカスタマイズされたOSである，SpeecysOSというOSを内蔵している．主な仕様を表3.1に示し，またサーボについては次節で詳細を述べる．



(a)正面　　(b)背面

図 3.1　実験に使用したロボット――SPC-001

表 3.1　SPC-001 の主な仕様

関節可動部自由度	頭	2軸
	腕	4軸X2本
	上半身回転	1軸
	脚部	6軸×2本
	合計	23自由度
頭部インターフェース	自由度	2軸
	ＬＥＤ（眼部分）	3色X2セット
	35万画素ＣＭＯＳ カラーカメラ	1個
	マイク入力	2個
	音声出力用スピーカー	1個
外部接続用 汎用インターフェース	無線LAN標準装備　（IEEE802.11b準拠）	
センサー	サーボ内蔵	サーボ個数分
	ジャイロセンサー	1軸　（2軸、3軸オプション）
	3軸Gセンサー	1セット
	CMOSカラーカメラ	1個
	マイク	2個（モノラルx2）
バッテリー	ニッケル水素専用電池9.6V　2000 mA	
寸法／重量	約５０cm／３．７Kg	
消費電力	約1.7A（ひざを曲げて静立時） 約4A前後（Speecys_Dancing時）	
動作時間	約15～30分（当社規定による測定）	
充電時間	約1～2.5時間（充電電流による）	
フレーム	アルミ合金製　アルマイト処理　3色	

3.2　サーボ：RS601CR

前節において述べた SPC-001 において使用されているサーボは，RS601CR と言いロボット用に開発されたサーボである．（図 3.2）

このサーボは，サーボの角度や負荷，温度，移動経過時間などサーボの状態を検出する事が可能である．また，移動後の角度と移動時間を指定する事によりロボットを動作させる役割を果たす．主な仕様を表 3.2 に示す．また，各サーボには ID 番号が振り分けられており，それぞれのサーボのロボット上の位置は図 3.3 のとおりである．

表 3.2　RS601CR サーボの主な仕様

寸法	59.0×26.0×47.1mm
重量	93g（軸、アルミホーンなしの状態で）
動作速度	0.17sec/60°　(9.6V)
可動範囲	240度
出力トルク	21kg･cm (9.6V)
電源電圧	9.6V
制御方式	RS485（127個まで接続可能）
通信速度	最大1.3Mbps
通信方式	双方向、コマンド式、半2重
センサー情報	角度、トルク、温度など
その他	移動経過時間の測定が可能.

図 3.2　RS601CR サーボ

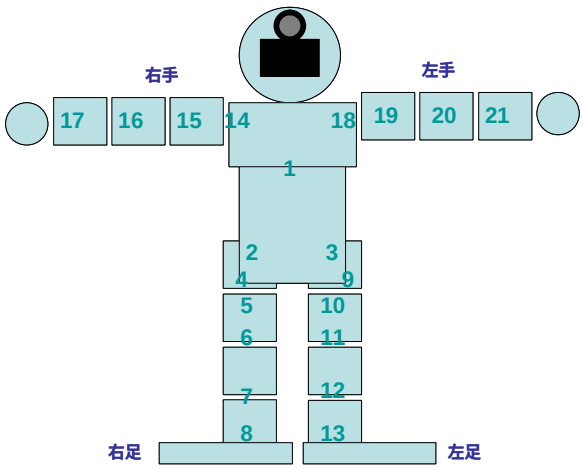


図 3.3　SPC-001 における各サーボの位置

3.3 まとめ

本章では，実ロボットの実験で使用するロボット，SPC-001 について説明し，本研究においてロボットのデータ取得の際に用い，またロボットの動作の要となったサーボについて述べた.

本研究における実験で用いられるロボットは,以後全てこの SPC-001 を指すものである.

第 4 章　予備実験：ロボットへの強化学習の適用

ロボットを用いた本実験の前に，予備実験として実ロボットに強化学習を適用し行動を学習させる実験を行った．

本章ではその予備実験について述べる．

4.1　実験目標

実ロボットにいくつかの行動と状態を与え，状態に応じた行動選択の学習を行わせる．

人間がロボットの腕や足などの部位に触れ，触れられた部分に応じてロボットが行動を行う．その行動が人間の望むものと一致していれば正の報酬・異なっていれば負の報酬を与え，触れられた部位に応じてどの行動が適しているか学習を行わせる．

学習終了後，ロボットが触れられた部位に応じ『人間の求める動き』ができるようになる事を目標とした．

4.2　実験方法

この節では，実ロボットの行動学習における実験方法及びその設定について述べる．

4.2.1　実ロボットの行動学習への強化学習法の適用

実ロボットに８つの状態と８つの行動を定義し，各状態に応じた行動を学習させる実験を行った．各状態及び行動を表 4.1 に示す．

人間がロボットの手足等に触れる事によって負荷を与え，サーボにかかる負荷の入力パターンによって状態を決定する．負荷のかかっているサーボを”1”／負荷のかかっていないサーボを”0”とした２進数でロボットが状態を認識し行動を行う．その行動を人間が評価し，状態に応じた行動であれば正の報酬／異なれば負の報酬を与える．得られた報酬を基にしてロボットはとった行動の良し悪しを学習する．

ロボットの行動学習の一連の流れを図 4.1 に，ロボットと環境の相互作用の概念図を図 4.2 に示す．ロボットが与えられた負荷による状態を検知し，その状態にあった行動を softmax 法と ε -greedy 法を組み合わせた行動選択手法によって選択する．そこでとった行動を人間が評価し良し悪しによって正か負の報酬を与える．その報酬を基に Q 学習法によりロボットはその状態でその行動をとることの価値，行動価値の更新を行う．

表 4.1　ロボットの状態と行動の定義.

状態		行動	
S0	右手に前後方向の負荷がかかる	A0	右手を上下に動かす
S1	左手に前後方向の負荷がかかる	A1	左手を上下に動かす
S2	右手に左右方向の負荷がかかる	A2	右手を左右に動かす
S3	左手に左右方向の負荷がかかる	A3	左手を左右に動かす
S4	右足に前後方向の負荷がかかる	A4	右足を前後に動かす
S5	左足に前後方向の負荷がかかる	A5	左足を前後に動かす
S6	右足に左右方向の負荷がかかる	A6	右足を上げて足首を動かす
S7	左足に左右方向の負荷がかかる	A7	左足を上げて足首を動かす

学習回数分繰り返す:
　　ロボットに負荷を与える.
　　ロボットが状態を認識.
　　選択手法により行動を選択し実行.
　　人間が行動を評価し報酬を与える.
　　Q学習法により行動価値を更新.
学習回数に達したら終了.

図 4.1　ロボットの行動学習の流れ.

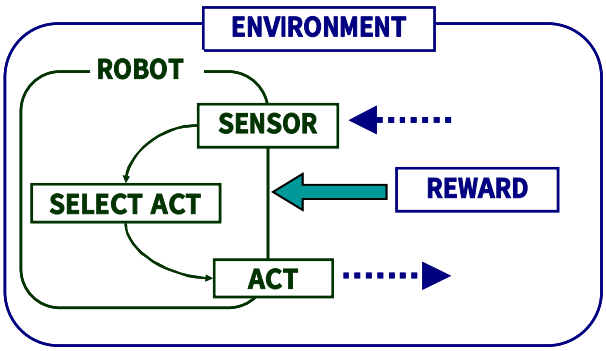


図 4.2　ロボットと環境の相互作用

4.2.2　使用した行動選択手法

今回の実験では，行動選択の際にε-greedy 法と softmax 法を混ぜた選択手法を用いた．

ε-greedy 法の欠点として，ほとんど最悪と思われる行動を選択する可能性とほとんど最適行動に近い良い行動を選択する可能性が同じくらいに高いという事が挙げられる．
softmax 法においては，Q 値に大きな差が無いと，行動選択確率に差が出ないという事がある．これらの欠点を考え，２つの手法を組み合わせε-greedy 法でのεの値を高めに設定し，(1-ε)の確率でε-greedy 法，確率εで softmax 法での行動選択が行われるという行動選択手法を考え用いた．

4.2.3　実験設定

表 4.1 に示した状態と行動を用い，実ロボットの行動学習の実験を行った．

人間がロボットに触れることで負荷を与え，その入力パターンをロボットが状態として認識し自らの持つ行動価値を基に行動を選択し実行する．その行動の良し悪しによって評価を行い，そこで得られた報酬を基にロボットは自分のとった行動の良し悪しを学習する．

人間が行動の評価を行う際の評価基準は，表 4.1 に示す状態の番号とロボットがとった行動の番号が一致したものであれば正の報酬・異なっていれば負の報酬を与えるものとした．その基準に従い，ロボットの背面についている A／B ボタンを用いて各状態における行動の評価を行った．正の報酬を与える場合は A ボタン・負の報酬を与える場合は B ボタンを割り当てた．

ロボットの行動選択手法としてはε-greedy 法と softmax 法を混ぜたものを用い，学習手法には Q 学習法を用いた．

以下に今回の実験における報酬及び各種パラメータの設定を示す．

報酬の設定：

正の報酬：　+10

負の報酬：　0

パラメータ設定：

α：　0.1

γ：　0

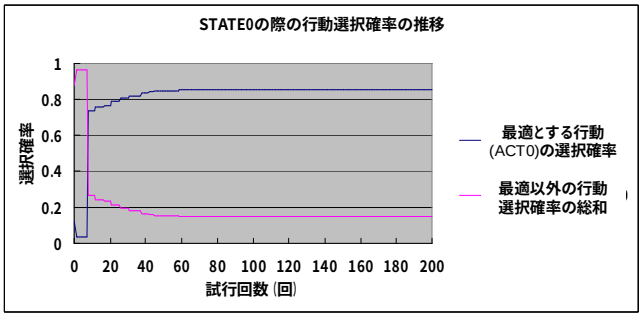
ε：　0.3

τ：　0.1

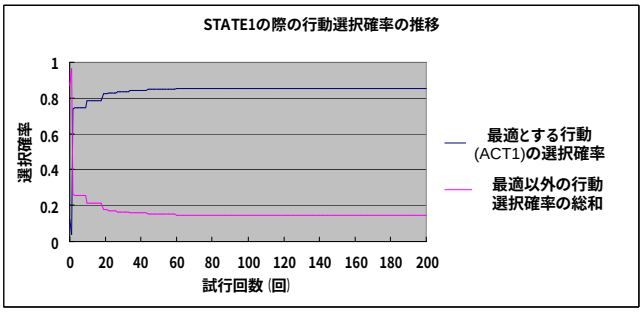
以上の設定のもと，状態の入力から行動評価・学習までの流れを各状態について 200 回繰り返し，学習させる実験を行った．

4.3　実験結果

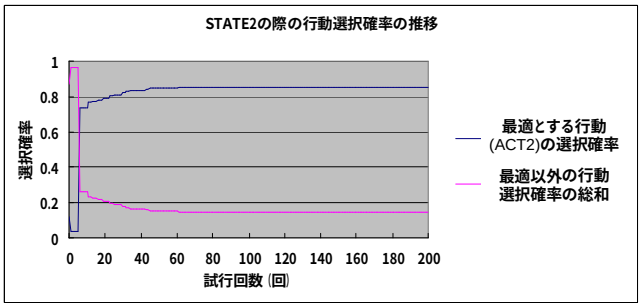
実ロボットに強化学習を適用し行動学習を行わせた実験結果を図 4.3 及び図 4.4 に示す．各状態において最適とする行動の選択確率の推移及び最適とするもの以外の行動の選択確率の総和の推移のデータを各試行回数において採取しグラフ化した(図 4.3)．また，学習終了後の各状態における各行動の価値（Q 値）を図 4.4 に示す．



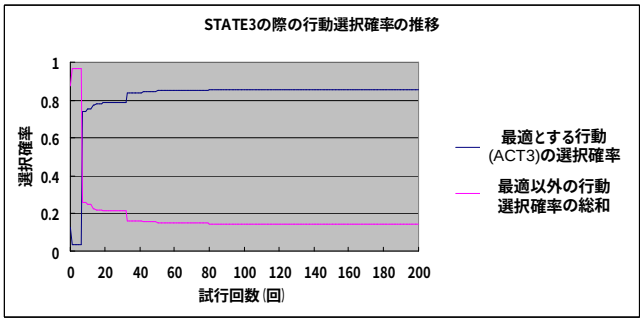
(a) STATE0 で最適とする行動の選択確率及びそれ以外の行動の選択確率の推移



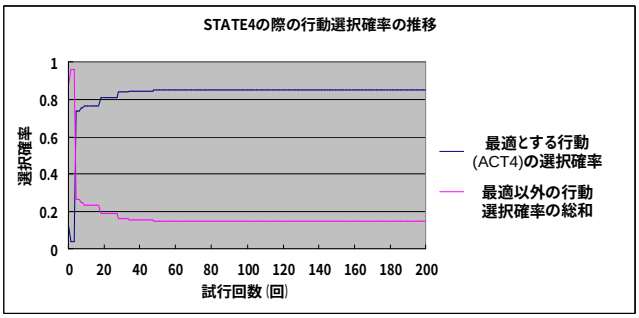
(b) STATE1 で最適とする行動の選択確率及びそれ以外の行動の選択確率の推移



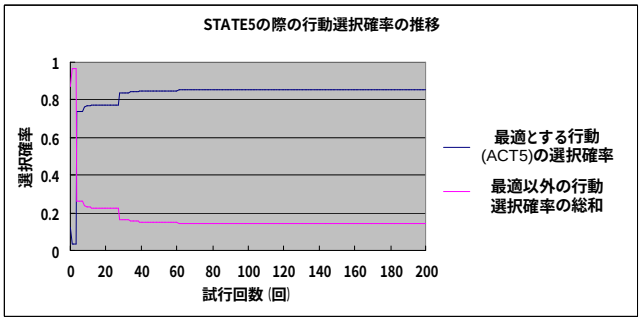
(c) STATE2 で最適とする行動の選択確率及びそれ以外の行動の選択確率の推移



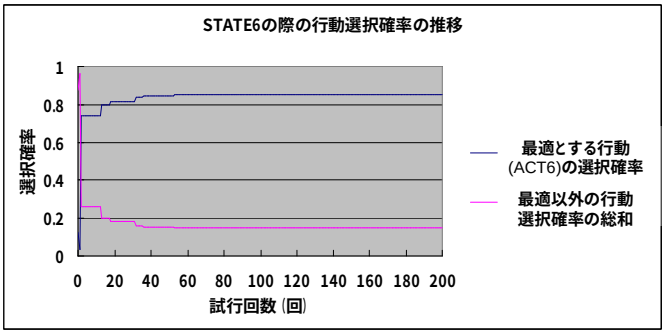
(d) STATE3 で最適とする行動の選択確率及びそれ以外の行動の選択確率の推移



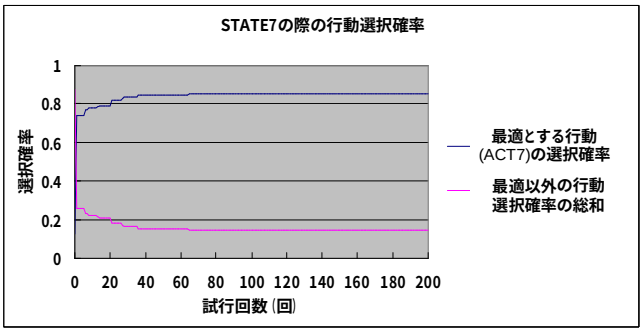
(e) STATE 4 で最適とする行動の選択確率及びそれ以外の行動の選択確率の推移



(f) STATE 5 で最適とする行動の選択確率及びそれ以外の行動の選択確率の推移



(g) STATE 6 で最適とする行動の選択確率及びそれ以外の行動の選択確率の推移



(h) STATE 7 で最適とする行動の選択確率及びそれ以外の行動の選択確率の推移

図 4.3　各状態において最適とする行動とそれ以外の行動の選択確率の推移

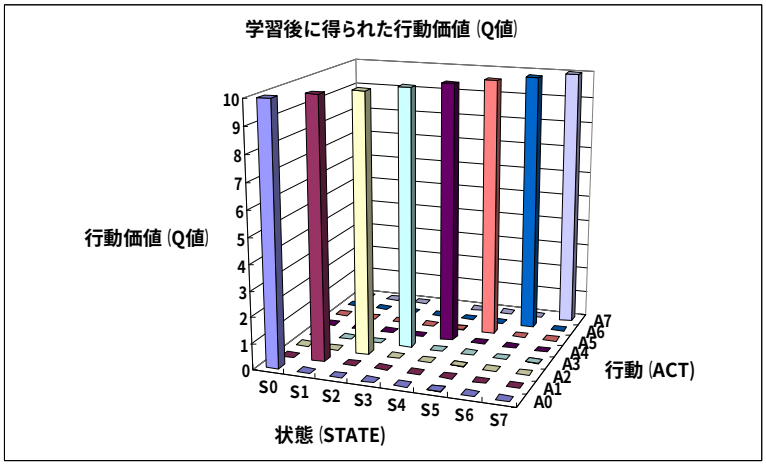


図 4.4　学習後の，各状態における各行動の行動価値（Q 値）

4.4　考察

行動選択確率の推移のグラフ（図 4.3）より，最適行動を選択する確率が試行回数の増加とともに高くなっていることがわかる．また，最適行動の選択確率が 90%付近で収束している事がわかる．また各状態において 200 回の学習をした後に得られた，各状態における行動価値のグラフ（図 4.4）より，各状態において状態の番号と行動の番号が一致している部分の行動価値，すなわち今回の実験で各状態において最適としている行動の価値が最も高くなっている．

これらの結果より，ロボットは最適行動を学習し，行動を選択するようになったと言える．

4.5　まとめ

本章では，強化学習により実ロボットに行動選択の学習をさせる実験を行った．

学習方法として Q 学習法を用い，行動選択手法として ε-greedy 法と softmax 法を組み合わせたものを用いた．それらの手法でロボットの各状態（表 4.1）において 200 回試行を行い，行動選択の学習が行われる事の確認を行った．その結果，4.4 節に示す実験結果より学習が行われた事を確認した．

また今回実験を行った中で，ロボットが状態を検出し一度行動を開始すると，途中で腕がひっかかるなど無理な動きになっても行動を続けようとするという事があった．それはそのまま動作を続けるとサーボの破損につながり，危険な事態である．そのためロボットの動作に逐次注意を払い，ひっかかる等した場合はそれを外してやる必要があった．

こういった，ロボットの動作中に起こる問題は本来人間が気にして逐一処理するものではなく，ロボットが自らその問題を処理する事が望まれる．動作中に障害物などの問題が起こった場合，ロボットが自律的にそれを避けたり動作を途中で止めたりして障害を回避する事が望ましいという事が改めて確認された．

第５章　痛覚を組み合わせたロボットのタスク学習

本章では，本論文で提案するシステムの概要及び手法を述べる．

5.1　手法の概要

本論文では，人間から与えられたタスクに対する行動学習と自己の安全確保のための行動学習を同時に行うシステムの構築を目標とする．その概念図を図 5.1 に示す．人が与えたタスクに対する行動学習でロボットが得る知識・学習結果は，人間が単発的に与えるタスクに特化したものであり与えられた目的に依存したものである．またその学習期間は，タスクを遂行する際のみであるので短期間の学習となる．これに対してロボットの安全確保などロボットにとっての本能とも言える学習は，自己の活動維持という永遠に持続されるであろう普遍的な目的のための学習である．そのためロボットが稼働し続ける限り継続し行われる．前者の学習を即時の目的（作業）のための学習，後者を普遍的な目的（生命維持）のための学習と考えると，これらの学習のレベルは異なるものである．これら異なるレベルの学習を統合し，同時に行わせることによって，ロボットが人から与えられたタスクのための行動・学習を行う中で，ロボット自身の安全確保のための行動・学習が行われる事になる．それぞれの学習が協調・競合しながらロボットの行動を生成する事で，自律的に危険を回避し与えられたタスクを遂行するシステムが実現されると考えた．

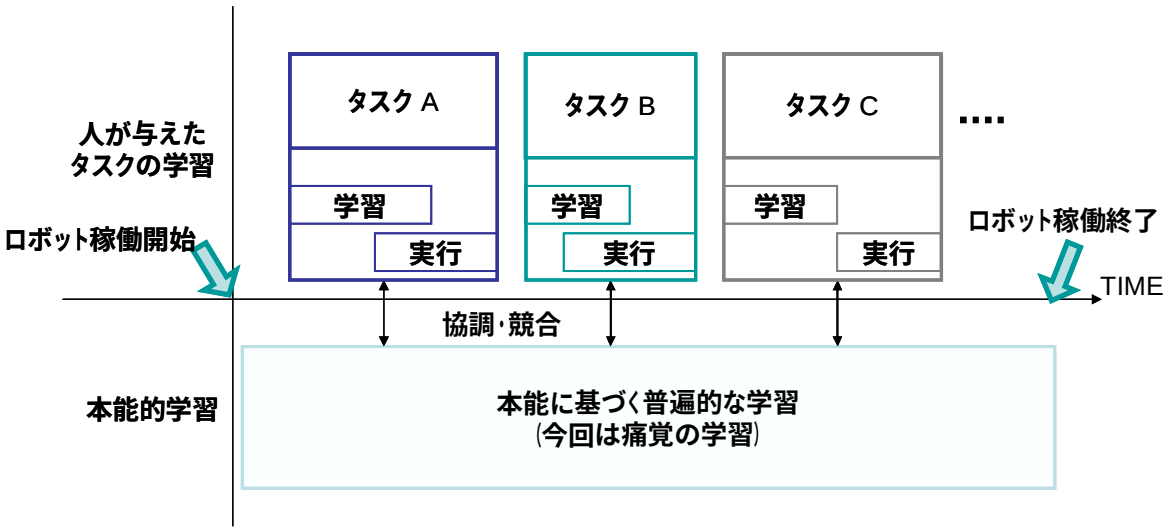


図 5.1　人から与えられたタスクの学習と痛覚の学習を行うシステムの概念図

5.2　痛覚を用いた本能の定義

ここで，ロボットの本能としてどのようなものを用いるかについて述べる．

人間同士で作業の依頼などをする場合，依頼者は作業の依頼をするのみで作業者に安全確保の方法などを教える事は無く，作業者は自分で自分の安全を確保しつつ作業を行う．これは，作業者が怪我をするなど身体の損傷につながる事態を，人間の本能の一部である「痛み」を通して危険であることを知り，回避するよう学習を行うからである．このような人間の本能における「痛み」をロボットにおいて実装し学習させる事で，自律的に安全確保を行うことのできるロボットが実現できると考えた．

ロボットにおいて，回避すべき危険な状態・ロボットの活動の継続ができなくなる状態には次のようなものがある．過負荷や過電流・発熱（異常温度）によるサーボ及びロボット内コンピュータの破損，外的衝撃による各部品の破損，電池不足による活動の停止などといったものである．これらの中で特に，人間が外から危険を判断し補助する事が困難であると思われるサーボの破損に着目し，ロボットにおける「痛み」の定義を行う．

サーボの破損に着目した場合，ロボットにとってそれを感知するためには負荷・電流・温度の検出が必要となる．そしてそれが異常なもの（過負荷や異常温度，過電流）であるかどうか判断する事によってサーボの破損を予期する事が可能となる．よって，ロボットにとっての「痛み」及び「痛覚」を以下のように定義する．

サーボにおいて検出した負荷・温度・電流をそれぞれL・T・Eとし，サーボの破損につながる過負荷・異常温度・過電流をそれぞれL_c・T_c・E_cとする．また，痛みを P とし，痛みのある状態を 1，無い状態を 0 で表すとすると，

$$L \geq L_c \cup T \geq T_c \cup E \geq E_c \rightarrow P = 1 \qquad (5.1)$$

で痛みのある状態が表される．これを「痛覚」と定義し以後用いる．

本論文では，この「痛覚」に基づいた行動学習と人から与えられるタスクに関する行動学習を同時に行う事により，自らの安全の確保を自律的に行いながらタスクの遂行をするロボットシステムを作成する．

5.3　人から与えられるタスク学習との統合

痛覚の学習と人からのタスクの学習を同時に行うには，それぞれの学習機構を統合する必要がある．その際に問題となってくるのが，行動の衝突である．これは，２つの学習においてそれぞれ行動が存在するため各学習が同じ部位対し，同時に・あるいは一方の行動中に行動指令を与えた際に，どちらの行動を優先するかという問題である．

異なるレベルの学習を統合するにあたって，このような行動の衝突が起こった場合に，どういった行動を優先させるかという指標を考え統合を行う必要がある．

5.4 まとめ

本章では，ロボットが人から与えられたタスク学習とロボットの本能に基づく学習を同時に行うという概念を示した．またロボットにおける本能として，ロボットにとっての「痛み」を定義しそれを用いた痛覚の学習を提案した．さらに，本能の学習とタスク学習という異なるレベルの学習の統合における問題点である行動の衝突について説明し，統合の際に留意すべき点などについて述べた．

第６章　実験：痛覚を用いた実ロボットの行動学習

前章で提案したシステムを実ロボットに実装し，人から与えられたタスクに関する行動の学習と痛覚を用いた行動学習を同時に実ロボットに行わせる実験を行った．本章ではこの実験について述べる．

6.1 実験目標

予備実験で行った行動学習に痛覚の学習を組み合わせ，人が与えるタスク学習（行動の学習）とロボットの本能的な学習（ロボットの活動維持を目的）を同時に行う実験を行った．これら２種類の学習を同時に行う事によって，ロボットの行動中に腕がひっかかる等の危険な状態が発生した際，ロボットによる自律的な回避行動が実現される事を目標とし実験を行った．なお，今回の実験では簡単のためにロボットの右腕にのみ着目し，実験を行った．そのため，ロボットの右肩に配置され，腕を前後に動かす際に用いられる 14 番のサーボ及び腕を左右に動かす際に用いられる 15 番のサーボにおいて痛覚の学習及び行動の学習が行われた．

6.2 実験方法

この節では，実ロボットによる痛覚を用いた行動学習の実験方法及びその設定・定義などについて述べる．

6.2.1 痛覚の定義

実験の際に用いたロボットにとっての「痛覚」の定義について述べる．

今回の実験では，ロボットの故障原因の一つとしてサーボの損傷に着目し，それを引き起こす過負荷・過電流・発熱（異常温度）のうち特に外部からの刺激に関わりの深い「過負荷」を今回の実験における「痛み」として用いた．よって，今回の実験における「痛み」は以下のように定義する．

サーボにおいて検出した負荷を L とし，サーボの破損につながる過負荷をそれぞれ L_c とする．また，痛みを P とし，痛みのある状態を 1，無い状態を 0 で表すとすると，

$$L \geq L_c \rightarrow P = 1 \qquad (6.1)$$

$$L < L_c \rightarrow P = 0 \qquad (6.2)$$

で痛みの有無が表される．ここでの過負荷は，ロボットの通常動作時にかかる負荷よりも大きな負荷であると定義し，用いている．

また，この「痛み」を感知する機構を，ロボットにとっての「痛覚」と定義し本実験において用いる．

6.2.2　過負荷の設定

「痛み」の判断基準として用いるロボットの過負荷の設定について説明する．

本実験では，ロボットの通常動作時にサーボにかかる負荷より大きい負荷を過負荷として定義する．通常動作時の負荷と異常な負荷及び過負荷とを区別するために，以下の予備実験を行った．今回実験に用いたサーボ 14 番においてロボットの通常動作時の負荷データを取り，その平均と分散を求め正規分布化した後，サーボにかかる負荷（サーボ状態）の「正常」「異常」の閾値，及び「過負荷」の閾値を計算し求めた．

通常動作としては，行動学習の際にサーボ 14 番が行う，右腕を上下に動かすという行動を用いて負荷データを取得した．なお，サーボ 15 番については装着角度が 14 番と異なるのみで，行う行動自体は同質であり装着場所もほぼ同じであるため，14 番の測定で得られた結果を 15 番の学習にも適用した．

ロボットが動作している間のサーボの負荷データを測定し記録することを 50 回行った．各回において約 1000 個の負荷測定データが得られたが，その中から最大値を求め動作中のサーボにかかる最大負荷の分布データが得られた．その分布を図 6.1 に示す．

前述の最大値の分布データから平均（以下μ）・分散（以下σ）の値を求めた．自然界の事象に存在するノイズは一般に正規分布に従う事から，得られたμ・σを基にデータを正規分布化した（図 6.2）．

正規分布において，「事実上の全て」の意味で用いられるのが，３σ範囲という考え方である．区間[μ-3σ，μ+3σ]の範囲に確率変数の事実上全てが入るというもので，３σ範囲の外へはずれる確率は千に三つとも言われる[2]．したがって，今回の実験において，得られた正規分布の３σ範囲から外れた値をサーボ状態の「正常」「異常」の閾値とした．その値はμ＋３σを超えた 79.8772 となった．また過負荷については，3σより大きく正常動作時にはまず得られないであろう負荷と考え，μ+5σを過負荷の閾値とし，その値は 89.662 となった．図 6.3 に得られたμ・σ・μ＋３σ・μ+5σの値を示す．

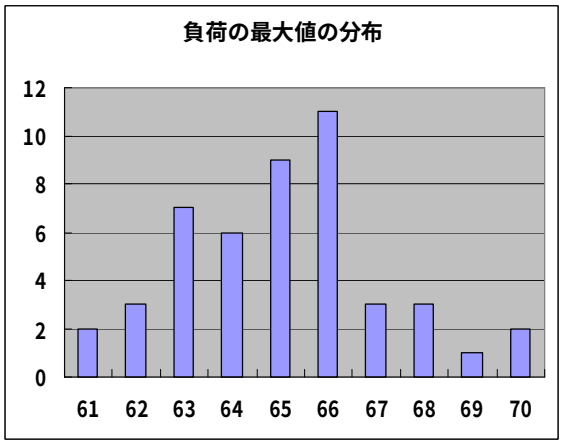


図 6.1：　動作中のサーボ負荷最大値の分布

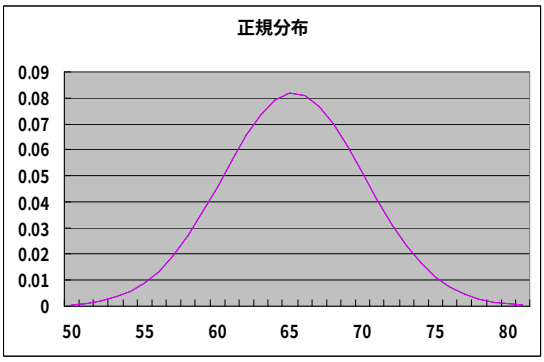


図 6.2：サーボ最大負荷の正規分布

平均 μ	65.26
分散 σ	4.8724
μ+3σ	79.8772
μ+5σ	89.622

図 6.3：平均・分散及び過負荷の閾値

6.2.3　痛覚と行動の学習への強化学習の適用

実ロボットに，行動学習について表 6.1 に示す状態及び行動，痛覚の学習について表 6.2 に示す状態及び行動を定義し，それぞれについての学習及び行動を同時に行う実験を行った．

行動学習における状態は，４章で述べた予備実験における状態と同様の方法でロボットに認識されるものであり，行動と報酬の与え方も同様である．行動選択手法には softmax 法を用いた．

痛覚の学習についてのロボットの状態は，6.2.2 で定義したサーボにかかる負荷が「正常」か「異常」かで状態の区分を行った．ロボットの行動としてはサーボの位置を初期位置に戻すという行動（回避行動）と，特別な行動指令を出さないという行動を設定した．痛覚の学習での各状態において，ロボットが選択する行動に対する報酬の与え方は次のように行った．学習促進のため，本来のタスク遂行の妨げとなる，サーボを初期位置に戻す行動には小さな負の報酬，何もせず行っているタスクを遂行させる行動には 0 を与えるよう設定した．これは「痛み」を感じたか否かに依存せず，表 6.2 におけるどちらの状態においても与えられる報酬である．

表 6.1　行動学習についてのロボットの状態と行動の定義

状態		行動	
S0	右手に前後方向の負荷がかかる	A0	右手を上下に動かす
S1	右手に左右方向の負荷がかかる	A1	右手を左右に動かす

表 6.2　痛覚の学習についてのロボットの状態と行動の定義

状態		行動	
S'0	サーボにかかる負荷がμ+3σ以下である	A'0	サーボを初期位置に戻す(回避行動)
S'1	サーボにかかる負荷がμ+3σ以上である	A'1	行動指令なし

また，ロボットには本来製作時から，ある一定以上の負荷がかかるとサーボのトルクをオフにするリミッターが実装されている場合が多い．しかし，本来のリミッターの発動する負荷は本当に危険な負荷である事が考えられる上，実験において何度もリミッターを発動させた場合，一度の実験でサーボが破損する危険性がある．そのため，リミッターの役割をする擬似的なハードウェアリミッターを実装した．実装した擬似リミッターは，サーボに，3σ範囲をはるかに超えるμ+5σ以上の負荷がかかり，それを 10 回連続で検知した場合に発動し，そのサーボのトルクをオフにするというものである．これが発動した場合，ロボットの破損にかなり近い行動を行ったと考えられるので，痛覚の学習機構に大きな負の報酬を与えるよう設定した．

図 6.4 に痛覚の学習と行動の学習を行う流れを示す．このように並列的にロボットに学習を行わせる実験を行った．

6.2.4　痛覚の学習と行動学習の統合

痛覚の学習と行動の学習を統合するにあたって，２つの学習においてそれぞれ行動が存在するため，同時に行動を行おうとした場合及び一方が行動を行っている場合に他方の学習で新たな行動を行おうとした場合において，行動の衝突が生じる．

こうした衝突の際にどちらの行動を優先するかについて，今回の実験ではそれぞれの学習において選択された行動の選択確率を、行動優先度として用いた．これは，行動選択確率が行動価値に基づいて得られるため，選択確率が大きいものほど，その学習の中での行動価値が相対的に大きいものであると判断されるためである．

これによって，ロボットが何らかの行動を実行中であっても，優先度の高い行動の指令がサーボに与えられた場合そちらが優先されるようになる．

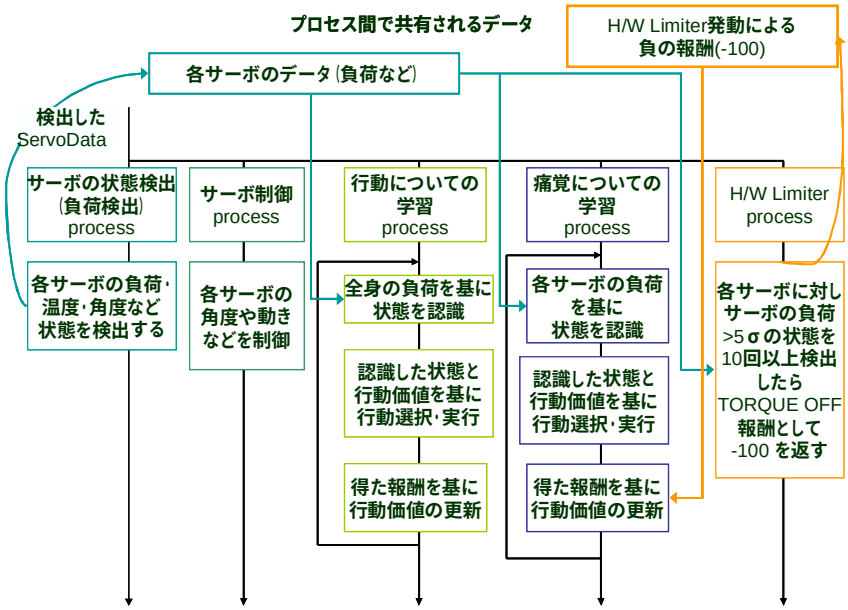


図 6.4　痛覚の学習と行動の学習を行う流れ

6.2.5　実験設定

表 6.1 及び表 6.2 に示した状態と行動を用い, 実ロボットについて痛覚と行動の学習の実験を行った．行動の学習における学習の流れ及び痛覚の学習における学習の流れ，そして各学習におけるパラメータ設定について述べる．

行動の学習については 4 章での予備実験と同様，人間がロボットに触れることで負荷を与え，その入力パターンをロボットが状態として認識し自らの持つ行動価値を基に行動を選択し実行する．人間がその行動に対し評価を行い，そこで得られた報酬を基にロボットが自分のとった行動の良し悪しを学習するというものである．

人間がロボットのとった行動の評価を行う際の評価基準は以下のように,

状態，行動の組が　(S0, A0), (S1, A1) ： 正の報酬；

(S0, A1), (S1, A0) ： 負の報酬；

と設定し, 表 6.1 に示す状態として与えられた負荷の方向と行動の際の動作方向が一致したものであれば正の報酬・異なっていれば負の報酬を与えるものとした．報酬はロボットの背面についている A/B ボタンを用いて与えられる．正の報酬を与える場合は A ボタン・負の報酬を与える場合は B ボタンを割り当て，各行動の評価を行った．

ロボットの行動選択手法としては softmax 法を用い，学習手法（行動価値の更新方法）として Q 学習を用いた．

痛覚の学習については，6.2.2 で述べたサーボ状態が「正常」か「異常」かをロボットは状態として認識し，自らの持つ行動価値を基に行動を選択し実行する．その行動に対し，各行動について定められた報酬値と擬似リミッターによる報酬を足し合わせ，報酬として得る．それを基にロボットが，その状態における自分の行動の良し悪しを判断し学習を行う．

各行動についての評価は，学習促進のために定めた報酬と擬似リミッターの作動による負の報酬の和によってロボット自身により評価される．学習促進のための報酬とは，6.2.3 にて述べたように，何事も無い場合にサーボを初期位置に戻す行動をとることは作業の中断に繋がるため小さな負の報酬を与え，何もせずそのまま作業を続けるという行動には 0 を設定するというものである．擬似リミッターの作動による報酬とは，破損につながる行動としてリミッターの作動によって大きな負の報酬が与えられるというものである．リミッターの作動は人間における怪我と同様なものとして考えている．これらの報酬値の定義基づいて，ロボットが自らの行動を評価し学習を行う．

ロボットの行動選択手法としては softmax 法を用い，学習手法として Q 学習を用いた．

以下に各学習におけるパラメータの設定を示す．

・行動学習について

報酬の設定：

正の報酬：　+5

負の報酬：　-3

パラメータ設定：

α： 0.1

γ： 0

τ： 3

・痛覚の学習について

報酬の設定：

各行動について設定している報酬：

行っている行動の継続： 0

サーボを初期位置に戻す： -1

擬似リミッターの作動により与えられる報酬： -100

パラメータ設定：

α： 0.5

γ： 0

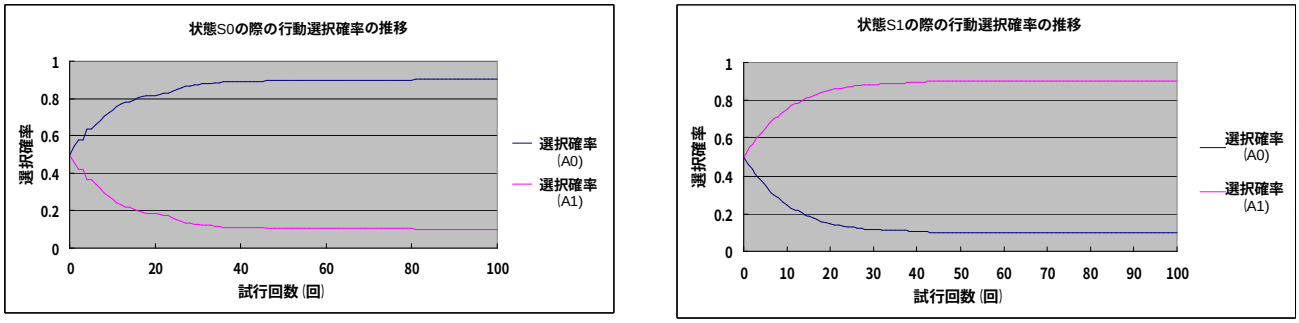
τ： 0.5

以上の設定のもと，痛覚と行動の学習の実験を行った．行動学習は各状態について 100 回ずつ行い，痛覚の学習はロボットの動作中，500msec に１回行われるようにした．

また，本実験はロボットの右腕にのみ適用したものであり，痛覚などは，ロボットの右肩に配置され腕の前後の動作に用いられる 14 番及び腕の左右の動作に用いられる 15 番のサーボにのみ実装したものである．

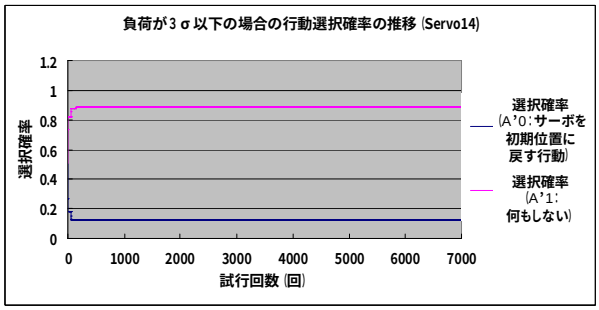
6.3 実験結果

行動の学習と痛覚の学習の実験を，実ロボットを用いて行った．その結果を図 6.5～図 6.7 に示す．行動の学習において，各状態における各行動選択確率の推移の様子を図 6.5 に示す．痛覚の学習において，14 番及び 15 番のサーボの各状態における各行動選択確率の推移の様子を図 6.6 に示す．またそれぞれの学習後に得られた各状態における各行動の価値（Q 値）を図 6.7 に示す．また，学習終了後に各状態の組でどのような行動がとられるか確認実験を行った結果を図 6.8 に示す．これは，各状態の組を 30 回ずつロボットに与え，その際に行われた行動の結果である．

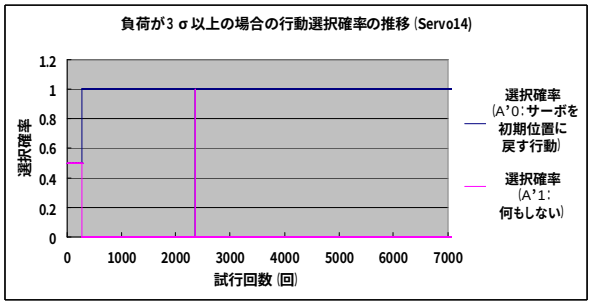


(a) 行動学習の状態 S0 での行動選択確率の推移　　(b)行動学習の状態 S1 での行動選択確率の推移

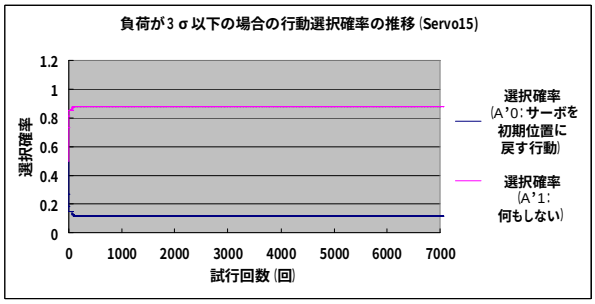
図 6.5　行動学習の各状態での各行動選択確率の推移．



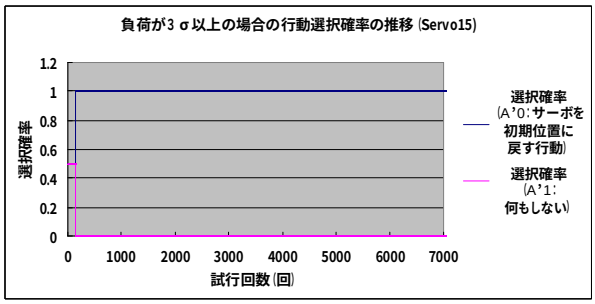
(a)　痛覚の学習の状態 S'0 での行動選択確率の推移（14 番）



(b) 痛覚の学習の状態 S'1 での行動選択確率の推移(14 番)

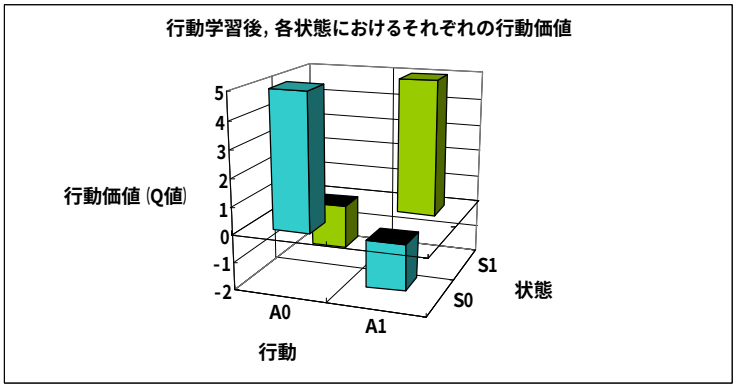


(c)痛覚の学習の状態 S'0 での行動選択確率の推移(15 番)

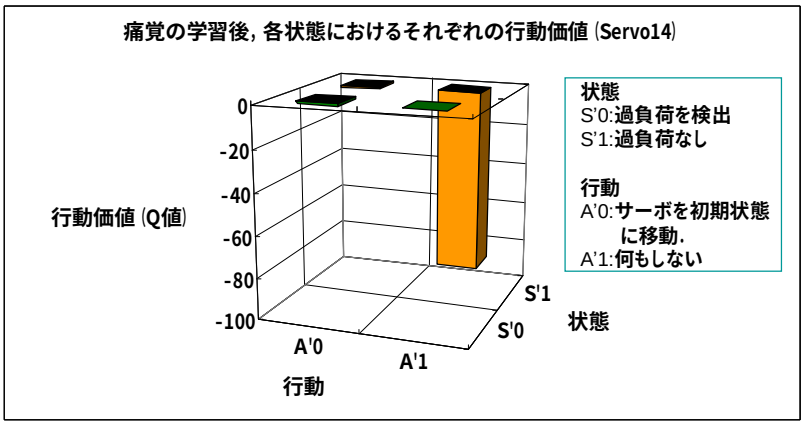


(d)痛覚の学習の状態 S'1 での行動選択確率の推移(15 番)

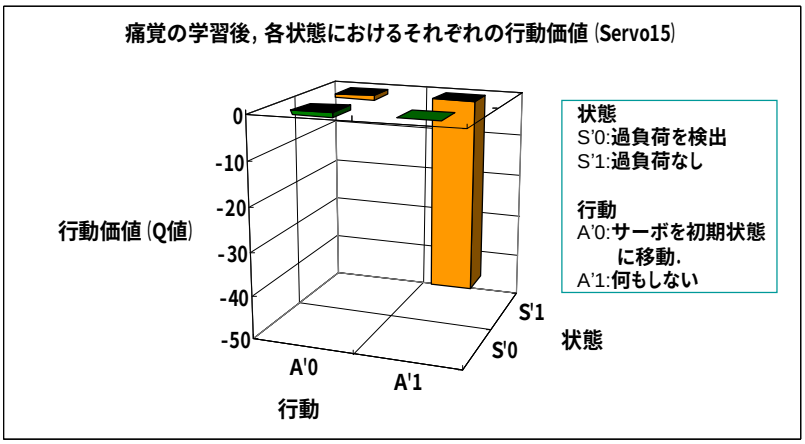
図 6.6　痛覚の学習での対象サーボの各状態における行動選択確率の推移



(a)　行動学習後の各状態における行動価値



(b)　痛覚の学習後，各状態における行動価値（Servo14）



(c)痛覚の学習後，各状態における行動価値（Servo15）

図 6.7　各学習後に得られた行動価値（Q 値）

	A'1(痛覚による行動指令なし)		
	A0（右腕を上下に動かす）	A1（右腕を左右に動かす）	A'0(回避行動)
S0（右腕に前後方向の負荷） && S'0（サーボ負荷が正常）	96.67%	3.33%	0
S0（右腕に前後方向の負荷） && S'1（サーボ負荷が異常）	0	0	100%
S0（右腕に左右方向の負荷） && S'0（サーボ負荷が正常）	6.67%	93.33%	0
S0（右腕に左右方向の負荷） && S'1（サーボ負荷が異常）	0	0	100%

図 6.8　学習終了後，各状態の組で実行される行動の確認結果

6.4　考察

行動の学習において，行動選択確率の推移のグラフ（図 6.5）より，最適行動を選択する確率が試行回数の増加とともに高くなっていることがわかる．このとき，最適行動の選択確率は 90%付近で収束している．また，各状態において 100 回行動の学習を行った後の行動価値のグラフ（図 6.7(a)）より，それぞれ状態番号と行動番号が一致する部分の行動価値値が同じ状態における他の行動の行動価値より高くなっている事が見られる．したがって，各状態において最適としている行動の価値が最も高くなっている事がわかる．これらの結果より，行動の学習においてロボットは最適行動を学習し，行動を選択するようになったと言える．

痛覚の学習において，行動選択確率の推移のグラフ（図 6.6）より，各サーボにおいて３σ以下の負荷である状態（S'0：サーボ状態が正常）では通常の痛覚の学習からは何の行動指令も行わないという行動（A'1）の行動価値が，３σ以上の負荷がかかった状態（S'1：サーボ状態が異常）ではサーボを初期位置に戻す行動(A'0：回避行動)の行動価値が高くなっている事がわかる．特に３σ以上の負荷がかかった場合においてはほとんど確定的な選択に近い値にまで選択確率が高くなっている．これは，リミッターが作動した場合の負の報酬が大きいためであると考えられる．また，図 6.6(b)において一瞬だけ通常行動とサーボを初期位置に戻す行動の選択確率が逆転しているが，これはサーボを初期位置に戻す行動を選択した際に行動が行われる前にリミッターが作動してしまったためである．そのため偶発的に起こったものであると考えられる．また，そのような偶発的な事が起こっても，学習を進めるうちに回避行動の選択確率が高いものへと戻っていることが分かる．また，各サーボについて学習後の各状態における行動価値（Q 値）のグラフ（図 6.7(b)(c)）より負荷が３σ以上かかった状態（S'1）において何もしないという行動をとった場合の行動価値が群を抜いて低いものとなっている．これは過負荷のかかった状態で行動を続けることに

よってリミッターがかかり大きな負の報酬を得てしまうということを学習したと言えるだろう.

これらの学習結果より，ロボットは行動の学習及び痛覚の学習を同時に行った事が確認できた.

また, 図 6.8 に示す学習終了後の各状態の組で実行される行動の確認結果より，人からのタスク学習における状態が与えられ・動作中の負荷が過負荷でない場合は，人からのタスク学習において各状態について最適としている行動が最も高い割合で行われる事が確認され，また人からのタスク学習における状態が与えられ・動作中の負荷が過負荷となった場合には回避行動が 100%の割合で行われる事が確認できた.

これより，ロボットの動作中に危険な状態（過負荷がかかった状態）になった場合，ロボットにおいて自己の安全確保のための回避行動がとられるようになった事が確認された.

6.5 まとめ

本章では，痛覚の学習と行動の学習を同時行うシステムをロボットに実装し実験を行い，その結果を示した.

学習方法としては Q 学習法を用い，行動選択手法には softmax 法を用いた．それらの手法でロボットの各状態（表 6.1）において 100 回試行を行い行動選択の学習が行われる事を確認した．また，行動選択学習の試行の中で並行して痛覚の学習を行い，痛覚の学習が行われた事を確認した．そして，学習終了後に各状態の組において実行される行動の確認を行った結果，ロボットが人からのタスクを遂行しながら，自己の安全確保を行う事が確認できた.

第７章　結論

7.1　まとめ

本論文では，人間から与えられたタスクに関する行動学習と自己の安全確保のための本能的な学習という異なるレベルにおける学習を同時に行うロボットシステムの構築を目標とした．その実現のために，ロボットにおける「痛み」をロボットの破損に繋がる事象として定義し，またその「痛み」を感知する一連の機構をロボットにおける「痛覚」として定義した．これを用い，「痛覚」に基づいた行動学習と人から与えられるタスクに関する行動学習を同時に行うロボットシステムを提案した．そしてその検証のため，提案したシステムを実ロボットに適用し，実験を行った．実験においてはサーボの破損につながる「過負荷」を用いてロボットにとっての「痛み」を定義し，痛覚に基づいた学習において使用した．また，人間が与えるタスクとして，人間の触れた部位を状態として検知し，その部位に応じた動きをロボットに学習させるという行動学習を用いた．これら２つの学習を組み合わせ，実験を行った結果，ロボットのサーボ状態（サーボにかかる負荷）が「正常」である場合は人から与えられたタスクに関する行動を行い，ロボットのサーボ状態が「異常」である場合はサーボを初期位置に戻す行動（回避行動）が行われるようになった．また，実験後に得られたそれぞれの学習における行動価値より，①人から与えられるタスクの学習においては各状態で人間が触れた部位に応じた行動の行動価値が他より高いものとなり②痛覚に基づいた行動学習においては，ロボットが痛みを感じる状態における回避行動の行動価値が他の行動と比べて大いに高いものとなった事が確認された．また，各学習の各状態において最適と思われる行動の選択確率も試行回数の増加に伴い高くなっている事が確認された．そして学習終了後，人間から与えられたタスクに関する学習での各状態と痛覚に基づく学習における各状態とを組み合わせ，それらの組でどのような行動が実行されるか確認したところ，かかる負荷が正常値の場合は人から与えられたタスク学習での行動が，負荷が異常値となった場合は回避行動が実行される事を確認した．

これらの結果より，痛覚に基づいた行動学習と人から与えられたタスクに関する行動学習が双方で行われ，またロボットの動作中に異常な負荷が検出された場合はロボットにおいて自律的に安全確保が行われる事が確認できた．

これより，提案した「痛覚を組み合わせた行動学習のシステム」が実際の行動学習の際に有用である事を示した．

7.2　これからの課題

本研究において，ロボットの「過負荷」の定義は人間がロボットのとる行動の，通常動作時の負荷を手動で計測し計算した結果によるものであった．今後，ロボットの行動を拡

張するにあたって，過負荷の計算も自動化しロボットによって定義／設定される事が望ましいと考えられる．

また，人からのタスクの学習とロボットの痛覚の学習の統合において，行動が衝突した場合の処理について今回はそれぞれの行動優先度を用いた．しかし，これはそれぞれの行動から得られる報酬に関わり無い指標であるため，本来の重要度を示すものではないと考えられる．したがって今後，それぞれの報酬を加味した統合方法を開発し用いる事により，それぞれの行動の重要度にあった統合が行えると考えられる．

謝辞

本論文を結ぶにあたり，日頃から様々な面で有益な御指導・御助言をいただきました倉重健太郎助手に深く感謝の意を表します．また多忙の中，お茶を淹れる事で研究をサポートしてくれた複雑数理モデル研究室吉田康志君に感謝の意を表します．

参考文献：

[1] R.S. Sutton and A.G. Barto. “ Reinforcement Learning: An Introduction. ” MIT Press, (1998). (邦訳 :“ 強化学習 ”, 三上, 皆川　訳, 森北出版, (2001)).

[2] 東京大学教養学部統計学教室. “ 基礎統計学Ⅰ　統計学入門 ”. 東京大学出版会 (2001).

[3] 安居院猛, 長橋宏, 高橋裕樹. “ ニューラルプログラム ”. 昭晃堂, (1993).

[4] 今田寛, 宮田洋, 賀集寛. “ 心理学の基礎 ”. 培風館, (2001).

[5] 森川幸人. “ マッチ箱の脳 (AI) ”. 新紀元社, (2000).

[6] 高橋泰岳, 浅田稔. “ 実ロボットによる行動学習のための状態空間の漸近的構成 ”, 日本ロボット学会, vol17, No.1, pp118-124(1999).

[7] 木村元, 山下透, 小林重信. “ 強化学習による４足ロボットの歩行動作獲得 ”, 電気学会　電子情報システム部門誌, vol.122-C, No.3, pp330-337(2002).