機械学習における
人間と機械の協調学習の研究

2013 年 9 月 30 日

指導教員　藪田哲郎　教授

横浜国立大学大学院工学府　システム統合工学専攻

機械システム工学コース

山科　亮太

Study on Cooperative Learning Method between Human Being and Machine for Machine Learning

by

Ryota YAMASHINA

Adviser: Prof. Tetsuro YABUTA

A Doctoral Dissertation Submitted to

the Specialization in Mechanical Engineering

in the Department of Systems Integration,

the Graduate School of Engineering of

Yokohama National University

September 30, 2013

目　次

第1章　序 論 1

1･1　概説 1
1･2　従来研究と本研究の目的 2
1･3　本論文の構成 6

第2章　ニューラルネットワーク概説と実験システム 7

2･1　概説 7
2･2　ニューラルネットワーク 7
2･2･1 ニューラルネットワークとは 7
2･2･2 ニューラルネットワークのモデル化 7
2･2･3 ネットワーク構造 9
2･2･4 ニューラルネットワークと学習 10
2･2･5 シグモイド関数 11
2･3　実験システム 14
2･3･1 実験装置の構成 14
2･3･2 実験装置の仕様 15
2･4　まとめ 22

第3章　学習結果Ⅰ　～人間と機械の協調学習～ 23

3･1　概説 23
3･2　実験内容 23
3･3　基本制御（非学習制御） 24
3･3･1 P 制御 24
3･3･2 PD 制御 27
3･3･3 PID 制御 29

3･4　学習制御 ･･･ 30
3･4･1　問題設定 ･･･ 30
3･4･2　人間と機械の接点：教示 ･･･ 30
3･5　学習の為のモデリング ･･･ 31
3･5･1　モデリング概要 ･･･ 31
3･5･2　システム同定 ･･･ 31
3･5･3　ARX モデル ･･･ 32
3･6　教示と汎化性 ･･･ 33
3･7　N.N.によるモデル/コントローラ繰り返し学習法 ･･･ 38
3･8　人間と機械の協調学習（役割分担） ･･･ 41
3･9　機械の学習過程の推察 ･･･ 44
3･10　教示の主観性 ･･･ 48
3･11　まとめ ･･･ 48

第4章　強化学習概説と実験システム ･･･ 49

4･1　概説 ･･･ 49
4･2　強化学習の概要 ･･･ 49
4･3　強化学習の特徴 ･･･ 50
4･4　強化学習の研究動向 ･･･ 52
4･5　強化学習の定式化 ･･･ 52
4･5･1　基本的な取り決め ･･･ 53
4･5･2　方策 ･･･ 53
4･5･3　収益 ･･･ 53
4･5･4　マルコフ性 ･･･ 55
4･5･5　マルコフ決定過程 ･･･ 56
4･5･6　価値関数 ･･･ 57
4･5･7　最適価値関数 ･･･ 57
4･6　TD 学習 ･･･ 58
4･6･1　TD 学習とは ･･･ 58
4･6･2　TD 学習の利点 ･･･ 60
4･6･3　Q-Learning ･･･ 60
4･6･4　Actor-Critic ･･･ 62
4･7　実験システム ･･･ 66
4･7･1　システム構成 ･･･ 66

4･7･2 実験装置の仕様 ····· 68
4･8 実験方法 ····· 74
4･8･1 Off-Line 学習プロセス ····· 74
4･8･2 状態パターン ····· 74
4･8･3 報酬の獲得 ····· 77
4･8･4 シミュレーション方法 ····· 78
4･8･5 実験方法 ····· 80
4･9 まとめ ····· 80

第5章 学習結果Ⅱ ～人間の報酬操作による機械学習支援～ ····· 81

5･1 概説 ····· 81
5･2 報酬情報 ····· 81
5･3 距離センサ報酬に基づく学習結果 ····· 83
5･3･1 前進移動距離の変遷（シミュレーション） ····· 83
5･3･2 行動価値関数の変遷 ····· 83
5･3･3 前進移動距離の変遷（実験） ····· 85
5･3･4 行動形態の変遷 ····· 85
5･4 報酬操作による学習改善 ····· 88
5･4･1 報酬操作の意図 ····· 88
5･4･2 報酬操作Ⅰ：段階報酬 ····· 89
5･4･3 報酬操作Ⅱ：強調報酬 ····· 96
5･5 まとめ ····· 100

第6章 機械学習における教示の新たな視点 ～教示の主観性･客観性～ ····· 101

6･1 概説 ····· 101
6･2 教示の主観性・客観性 ····· 101
6･3 問題設定 ····· 102
6･4 シミュレーションシステムの構成 ····· 103
6･4･1 Open Dynamics Engine (ODE) ····· 103
6･4･2 ODE の特徴 ····· 104
6･5 ODE の構成 ····· 104
6･5･1 オブジェクト ····· 104
6･5･2 シミュレーションフロー ····· 105

6･5･3 衝突検出 ····· 106
6･6 生物型ロボットモデルの作成 ····· 107
6･7 評価方法 ····· 109
6･8 まとめ ····· 109

第7章 学習結果Ⅲ ～主観報酬を通した人間の教示特性の理解～ ····· 110

7･1 概説 ····· 110
7･2 主観報酬情報 ····· 110
7･3 学習結果 ····· 111
7･4 主観報酬と客観報酬の学習結果比較 ····· 112
7･5 問題設定 ····· 114
7･6 人間の優れた教示能力Ⅰ：相対教示 ····· 115
7･6･1 仮説 ····· 115
7･6･2 検証 ····· 115
7･7 人間の優れた教示能力Ⅱ：教示分布 ····· 116
7･7･1 仮説 ····· 116
7･7･2 検証 ····· 117
7･7･3 考察 ····· 120
7･8 人間の優れた教示能力Ⅲ：形状認識 ····· 120
7･8･1 仮説 ····· 120
7･8･2 検証 ····· 121
7･8･3 考察 ····· 123
7･8･4 形状認識能力 ····· 123
7･9 機械学習における投影教示 ····· 129
7･10 まとめ ····· 130

第8章 結 論 ····· 131

第9章 今後の展開 ～機械学習における主観性の問題～ ····· 133

謝 辞 ····· 137

文 献 ····· 138

第1章
序 論

1-1 概説

人間が自然に行っている学習能力と同様の機能をコンピュータで実現しようとする技術・手法は機械学習と呼ばれ，これまで様々な手法が提案されている[(1)~(42)]．例えば人間の脳の構造をモデル化したニューラルネットワーク（N.N.）[(1)~(10)]，人類や動物の進化の過程をモデル化した遺伝的アルゴリズム（G.A.）[(11)~(22)]，生物の適応過程をモデル化した強化学習[(39)~(42)]などがある．これらの機械学習は様々な観点で分類・整理されるが，一般的な分類方法として「教師あり学習」と「教師なし学習」という分類がなされる．

両者の違いを人間の学習過程に置き換えて説明する．

例えば「速く走る」動作を生徒が獲得する場合を考える．教師が手の振り方，足の動かし方をお手本として見せて，生徒はそれを真似してお手本に近づくように練習して微調整を繰り返しながら，徐々に速く走れるようになる．このような学習者が教示者から正解動作を直接与えられる学習手法をソフトウエアにより再現したものを「教師あり学習」といい，代表的な学習アルゴリズムにはN.N.やサポートベクターマシン[(23)~(27)]などが挙げられる．

一方正解動作そのものの直接的な教示を受けなくても速く走れるようになることは可能である．例えば生徒は自分でいろいろと工夫しながら走ってみて，その結果をストップウォッチのタイムという形で与えられる．走ったタイムが良ければ，何が良かったのかを自分自身で理解し，悪ければ改善点を自分なりに考え，次の走法につなげる．これはストップウォッチのタイムという動作の良し悪しを示す指標は与えられるが，具体的な正解動作を直接的に与える教師がいないという意味において「教師なし学習」と呼ばれ，代表的な学習アルゴリズムとしては，G.A.や強化学習がある．

上述のように教師あり学習では機械へ“正解”が，教師なし学習では“ヒント”が与えられ，学習が進められる．このようなアルゴリズム的相違がある為，便宜的には大別され，それぞれの長所・短所を議論されることが多いが，両者の学習手法に機械学習としての本質的な違いは無い．なぜならば，いずれの学習手法も外界から何らかの有益な情報（教師あり学習で言えば手足の正解動作，教師なし学習で言えばストップウォッチのタイム指標）を受け，その情報を基に，設定された評価関数を最大化するようにパラメータ調整が行われる最適化問題に他ならない為である．それ故，これまでの機械学習の研究の多くは，「教師あり学習」「教師なし学習」問わず，最適化問題をいかに効率的に解くか？という問題に焦点が当てられてきた側面が強い[(12),(13),(24)]．

ここで，機械学習から見た，機械と人間の学習の本質的な違いについて述べる．

機械学習の限界は評価関数自体を機械が自ら学習することが出来ない点にある．

人間において，例えば最もわかりやすい評価関数として「自身の生命を維持する時間」がある．病気になれば病院に行き，猛スピードで走る車が近付けばその場から離れることで，人間は自身の生命を維持する時間を最大化しようとする．しかしひとたび自分の子供が生まれると，我が子の為に自らの命を投げ打ってでも，我が子を守ろうとする． このように一見最も揺るがないように思える「自身の生命を維持する時間」という評価関数ですら，あるきっかけで，優先度（評価関数の重み）が下がることがある．機械はこのような評価関数自体を柔軟に変化させることを自ら学習できるであろうか？答えはNoである．

このように機械が何かを学習する為には，人間が機械に対して評価関数と，その重みを与えなければならない．機械は評価関数を最大化することはできても，評価関数自体を学習することができないのである．

この問題は機械の「心」を扱った哲学的な問題を含む，機械学習の本質的な問題であるが，1959年にアーサー・サミュエルが「機械学習」という言葉を定義してから50年以上が経った今もなお，この問題に対する本質的なブレークスルーは起きていない．このため所謂「機械は人間を超えられるか？」という哲学的観点での問いかけに対してはNoと言うことになる．

機械で人間の学習能力と同様の機能を実現しようとする取り組みは大変意義深く，興味の尽きない問題であるから，この観点での機械学習の研究は今後も続けられていくであろう．しかし一方で，機械と人間の本質的な違いを正しく認識し，受け入れ，両者の得意・不得意な点をカバーし合う相互学習の関係，すなわち**機械学習における人間と機械の協調学習の枠組み**を構築することも非常に重要と考える．

機械と人間のインタラクションの問題は，**協調制御**という形で，過去に多くの研究例があるが[218],[219]，機械学習分野における，**協調学習**の側面でのインタラクションの研究は前例が非常に少ない．

本研究の目的は，機械と人間の違いを論ずることにあるが，単に対比して議論するだけではなく，両者の望ましい関係性を機械学習の枠組みの中で議論することにある．具体的には，機械学習における重要な特徴である「教示」に焦点を当て，特に教示の主観性の問題を中心に，人間と機械の違いを明らかにしていく．

1-2　従来の研究と本研究の目的

前述したように機械が望ましい動作を学習する為には，その良し悪しを示す外界からの何らかの情報が必要となる．以降学習に必要なこれらの提供情報全般を教示と呼ぶ．（“教示”は狭義では教師あり学習における「教師データ」のことをさすが，本論文内では人間が機械に教える情報全般の広義をさす．）

教示は人間が設計して機械に与えるものであるから，教示をどう与えるかの決定権は人間にある．ここで従来の機械学習の研究における，教示の与え方を述べる．

従来研究１　強化学習における教示（報酬）の与え方

強化学習は教師なし学習に分類され，機械の行動の良し悪しを示す“報酬”というスカラー量を教示することで学習が行われるが，従来研究の報酬の与え方を調査すると，以下の２つの観点で大別されていた[187]~[202].

(1) エピソードタスク型報酬 / 連続タスク型（即時型）報酬

一連のタスクにおいて，タスク終了時に報酬が与えられる問題を“エピソードタスク型問題”と言う．例えばオセロや将棋では，対局が終了した時点で報酬が与えられる．また迷路問題では，エージェントが試行錯誤の末，迷路をゴールした際に報酬が与えられる．これらの問題では報酬がタスク終了時のみに与えられるので，報酬の情報量は少ないが，知識，意図の介在が最小限である為，人間が想定しえない解を導き出す可能性が高く，強化学習の本来の利点が生かされた問題設定とも言える．

一方で，エピソードの途中で逐次報酬が与えられる問題を“連続タスク型問題”と言う．例えばロボットの前進行動獲得を例に取れば，一回一回の状態遷移毎，あるいは数回の状態遷移毎に報酬が与えられる．これらの問題ではゴールに至るまでの間に報酬という形で，複数回，知識，意図を与えるため，エピソードタスク型報酬による学習よりも，学習が速やかに進むなどのメリットを有する．

(2)　2 値報酬 / 多値報酬

報酬は与える値によっても分類することが出来る．○か×，Yes or No，成功 or 失敗で与えられるような報酬を 2 値報酬，距離センサなどの値に基づき連続的に与えられるような報酬を多値報酬と呼ぶ．例えばオセロの問題では，対局終了後勝ったか，負けたかのみが与えられる場合は，報酬は 2 値報酬となり，34 – 30 など具体的なスコアでその良し悪しの程度が与えられる場合は多値報酬となる．またロボットの前進行動の例で言えば，前進したか後退したかのみが与えられる場合は 2 値報酬となり，進んだ距離がセンサなどから連続的に与えられる場合は多値報酬となる．(1)と同様に 2 値報酬は情報量の少なさゆえに，学習速度は一般に遅いが想定しえない解を導き出す可能性が高く，逆もまた然りである．2 値報酬は多値報酬の一部であるとも考えられる為，本質的な分け方とは言い難いが，過去の研究を俯瞰すると報酬をシンプルに 2 値で設定する場合と，細かく多値で設定する場合に大きく二分されているため，便宜的に分類した．

以上のように，強化学習における教示（報酬）の分類は(1)(2)の通りであるが，これらの報酬の適切な与え方という観点での研究は宮崎ら[187]が理論的に，あるいは荒井ら[190]が実験的に検討を行い，これらの研究により強化学習における「報酬」の取り扱いの重要性と，設計指針が認識されている．

従来研究２　ニューラルネットワークにおける教示（教師データ）の与え方

ニューラルネットワークは一般には正解自体を教える教師あり学習に分類される．学習に用いるデータを“教師データ”と言い，教師データは人間が適切に設計して与えなければならない．

教師データは一般にはベクトル量で与えられる．例えばロボットアームの姿勢制御においては，望ましいモータの指令電圧に対する角度，角速度などの時系列データが[1],[3]，画像認識処理においては，正しい画像を認識する為の 2 次元あるいは 3 次元の空間データがベクトルで[4]与えられる．機械はこの教師データを基に，どのような入力に対してどのような出力をすればよいかという，入出力の望ましい関係を学習することになる．

従来研究 1,2 に示すように教示は

- 正解を与えるのか，ヒントを与えるのか？
- ベクトルデータなのか，スカラーデータなのか？
- 教示を与えるタイミング，値はどういったものか？

といった観点で分類されるが，上記の研究で与えられる教示はいずれも機械的なセンサの値を利用したものや，タスクの成否を表す〇か×といったものであり，これらは教示者の個々人の違いや，個人の心理的性質に依存しない，客観的な判断基準に基づき与えられるものが大部分を占めていた．

教示における主観性

主観とは客観の対義語であり，広辞苑によれば**「自分ひとりの考え方や感じ方」**である．

従来明確に分類がなされていない機械学習における教示分類の新たな観点として，教示の**客観性**・**主観性**が考えられる．人間の学習過程においては，教示は主観的教示，客観的教示の両側面を持っていると言える．例えば 1+1=2 であることは，どの教師も共通して生徒に教えるであろう．これは「自分ひとりの考え方や感じ方」を介入する余地が無い**客観的教示**である．一方で主観的な教示もある．例えば厳しい教師や優しい教師がいて，生徒の同じ行動に対しても，それを褒めたり叱ったりする．また同一の教師であっても，生徒の性格や成長過程に応じて適切に褒め方を変えたりすることもある．このような教示は「自分ひとりの考え方や感じ方」により異なる**主観的教示**であると言える．機械学習には，機械の自律性を強く求められている背景があるから，従来の機械学習の研究の大部分が，設計者の個性に依存しない客観的教示であったことは，当然とも言える．しかし，人間の学習における教示には，このような個々の人間や人間の心理的性質に依存した主観性の要素は強く，機械学習における教示の主観性の問題は非常に興味深い問題である．また主観性は機械では表現の難しい人間の特徴と考えられるので，人間の優れた能力，機械の優れた能力を対比して理解する上で重要な手がかりになると考える．

人間の主観的な評価指標に基づく機械学習は，過去にいくつか研究例があり，古くは進化的計算を用いたものが有名である．これらは対話型進化計算 IEC（Interactive Evolutionary Computation)と呼ばれる[210]~[217]．IEC の歴史は古く 1986 年に遡るが，初期は画像生成を中心とするアート分野への応用研究が行われてきた．その後感性情報処理などの研究をもとにした工学，教育などの実用的な研究が行われてきた．その他，室内インテリアのレイアウト，補聴器フィッティングなどの幅広い分野に展開された．対話型進化計算法における代

表的な手法はGAを用いたIGAである高木らの研究が有名である[212), 213)]．これらの研究は，人間の感性にあったモノ・空間を提供する為に，それらを構成するものにパラメータ（調整因子）を設け，パラメータ学習を人間の評価指標に基づき行うというものであり，最終的には評価関数の最適化問題に帰着する．高木らのIEC の研究は大変興味深く，個人の主観的な評価に基づき最適化学習が行われることで，個人に調和した学習結果が得られることを示しているが，人間と機械のインタラクションに積極的な焦点をあてられたものではなく「人間の優れた能力とは何か？」「機械の優れた能力とは何か？」に解を与えるものではない．

一方，動作を伴うロボットの学習に対し人間が主観的な教示を行い，ロボットの学習過程や学習結果の違いを考察した研究例は少ないが，いくつかの研究例がある[203)~209)]．Andrea ら[203), 204)]はロボットがケーキを焼くというタスクにおいて，“移動して必要な道具を集める”，“卵を割る”などの意思決定の優先順位を学習させるのに人間の主観的な評価を取り入れている．この研究では人間の適切な介入により機械学習が促進される結果が示されており大変興味深い．しかしロボットの意思決定という上位層レベルの問題にのみアプローチしており，検討が限定的である．一方廣川ら[205)~207)]は，下位層レベルの問題として倒立振子の振り上げ動作をタスクとして，人間が主観的に報酬を与えることにより，振り上げ動作の学習を誘導・促進できることを示している．この研究は大変興味深いものであるが，客観的な報酬と主観的な報酬の優劣などの詳細な分析は行っておらず，また被験者は一名のみであり，複数被験者間の報酬の違いによる学習結果の違いについても言及していない為，人間と機械の協調関係を体系的に理解するまでには至っていない．

以上従来研究調査から，機械学習の教示における主観性／客観性の問題を体系的に捉え，まとめられた研究論文は過去に例が無い．また「教示」という観点で，人間と機械の特徴の違いや，両者の望ましい協調学習の枠組みを構築した例も無い．

本研究の目的

本研究は人間が機械に与える教示と学習結果の関係を通じて，人間の持つ優れた能力，機械の持つ優れた能力を明らかにする．特に教示の主観性に着目し，機械には無い人間特有の優れた能力を明らかにする．

具体的には 2 つのロボットシステムに対して，2 つの異なる学習アルゴリズムを適用し，教示と学習の関係を多角的に議論する．はじめにフレキシブルアームロボットの制振問題をタスクとし，ロボットに自ら振動抑制をするニューロコントローラを獲得させる．機械がコントローラ学習を行う過程で，人間は機械へどのような教示ができるのか？また人間と機械の協調学習はどのような形態が望ましいのかを明らかにする．

次にイモムシ型ロボットの前進行動獲得をタスクとし，強化学習を用いてロボットに前進行動を自ら獲得させる．ここでは「報酬の主観性/客観性」の議論を展開し，人間がロボットに前進行動を学習させる上で，人間が主観的に行っている機械にはない優れた教示能力を明らかにする．また複数の被験者間の教示方法の違いを比較し，教示の主観性が与える機械学習への影響を議論する．

2011 年 3 月，東日本大震災が起き，震災の復旧復興に多くのロボットが活躍した．人間にしかできないことがあり，機械にしかできないことがある．しかしながら必ずしもそれが，人間が得意なこと，機械が得意なこととは一致しない．本研究の取り組みが，人間のスキル・個性のロボットへの投影という形で展開され，人間の持つ優れた能力を，機械にしかできない環境で実現できる未来を期待する．

1-3　本論文の構成

第1章　序論

本研究の背景と目的を述べる.

第2章　ニューラルネットワーク概説と実験システム

ニューラルネットワークアルゴリズムの概要と本研究で用いたフレキシブルアームロボットの実験システムについて記述する.

第3章　学習結果Ⅰ　～人間と機械の協調学習～

ニューラルネットワークによる振動制御の学習結果を示す．人間がロボットに与える各種教示と学習結果の関係を論じ，機械学習における人間と機械の協調の望ましいあり方を述べる.

第4章　強化学習概説と実験システム

強化学習アルゴリズムの概要と本研究で用いたイモムシ型ロボットの実験システムについて記述する.

第5章　学習結果Ⅱ　～人間の報酬操作による機械学習支援～

強化学習による前進行動獲得の学習結果を示す．報酬と言うシンプルな情報から，ロボットが前進行動を獲得する過程を述べると共に，人間が報酬を操作し学習に介入することでロボットの学習を促す結果を示す.

第6章　機械学習における新たな視点　～教示の主観性・客観性～

前章までの結果を受けて，機械学習における人間と機械の関係を整理すると共に，新たな視点である教示の主観性/客観性について，そのコンセプトと実験方法を述べる.

第7章　学習結果Ⅲ　～主観報酬を通した人間の教示特性の理解～

センサが与える客観的な報酬と，人間が与える主観的な報酬とで得られる前進行動獲得野結果を比較し，主観報酬による学習の優位性を示す．また人間はロボットの行動に対してどのような視点で教示を行っているのかを明らかにし，機械には無い人間の優れた教示能力を考察する.

第8章

本研究で得られた結論を記述する.

第9章

本研究で得られた結論に対して，今後の研究の方向性を述べると共に，機械学習における人間と機械の協調学習の問題及び，教示の主観性の問題が，今後どのような研究分野へ発展していくか，その展望を述べる.

第2章
ニューラルネットワーク概説と実験システム

2･1　概説

本章ではニューラルネットワークについての概説と実験システムについて記述する.

2･2　ニューラルネットワーク

2･2･1　ニューラルネットワークとは

ニューラルネットワークは人間の脳内の神経回路網を数学的にモデル化したもので，日本では情報処理の分野において，80 年代から音声認識や画像処理などに応用されてきた．現在も非線形な事象に対し柔軟性を持つアルゴリズムとして注目されている．ここではニューラルネットワークの数学的モデル化とネットワークの構造について触れる.

2･2･2　ニューラルネットワークのモデル化

人間の脳内ではニューロンと呼ばれる神経細胞が 140 億以上も存在し，一つのニューロンについて見れば，1 万程の他のニューロンと相互に結合しながら情報の受け渡しをしていると言われている．ニューロンは Fig.2-1 に示すように細胞体，入力部の樹状突起，出力部の軸策の 3 部から構成される. さらに軸策の先端部分は, 活動情報を伝えるシナプスと呼ばれる構造になっており，他のニューロンとシナプス結合をしている．シナプス結合の仕組みは，Fig.2-2 に示すように，他の細胞膜との間にシナプス間隙と呼ばれる微小な隙間が存在し，そこに化学物質が放出されて刺激（電位）が伝わるようになっている.

簡単に情報伝達の仕組みを段階化すると次のようにまとめられる.

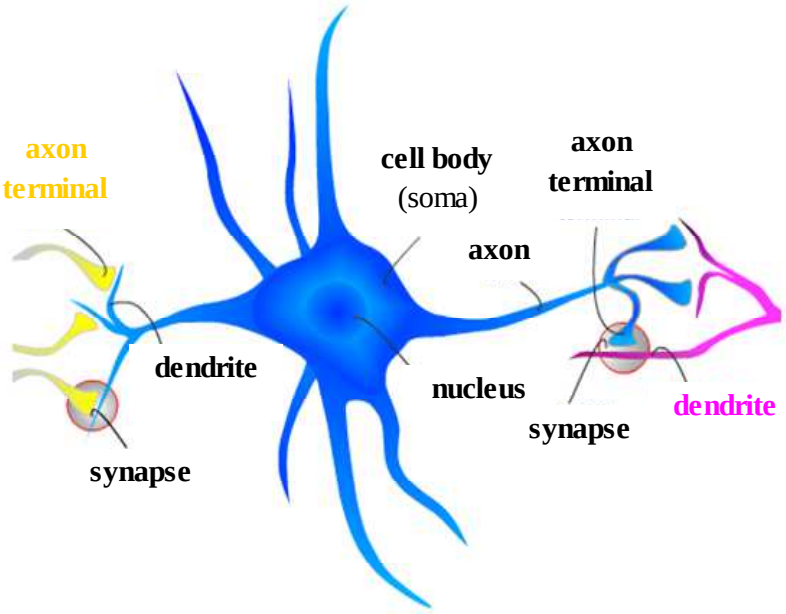


Fig.2-1 Neuron model

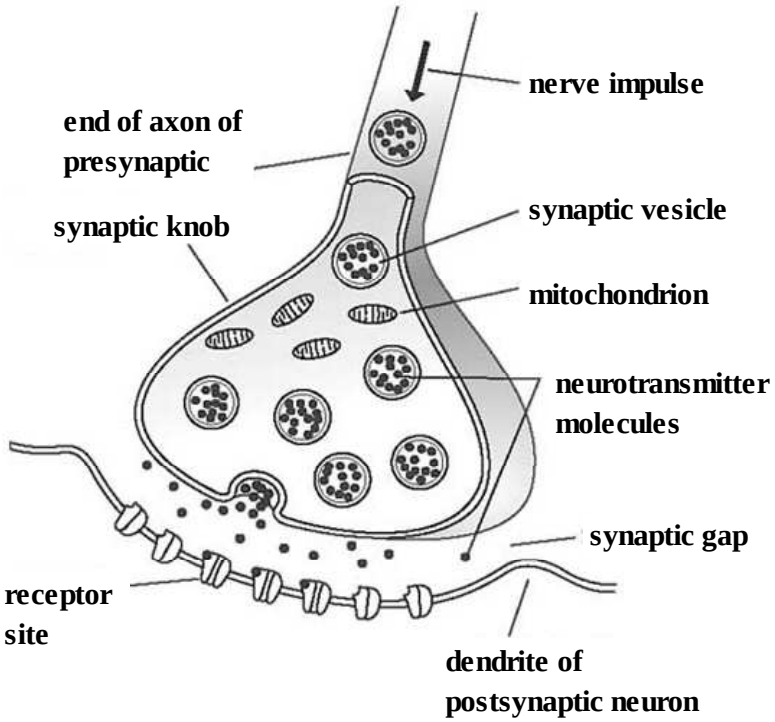


Fig.2-2 Neuron model

（1）他のニューロンのシナプスからの刺激（電位）の和が一定の閾値を超えると細胞体内部でスパイク状の活動電位が発生する．（活動電位の発生）

（2）軸策を伝わって活動電位がシナプスへ送られる．（神経インパルス）

（3）神経インパルスがシナプスに到達し神経伝達物質と呼ばれる化学物質をシナプス間隙（次のニューロンとの間隙）に放出する．

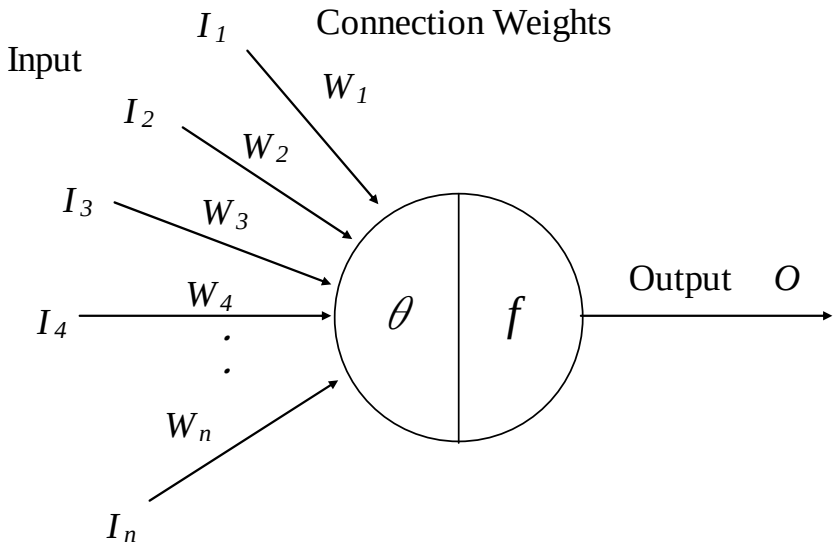


Fig.2-3 Neuron model

この構造を数学的にモデル化したものがニューラルネットワークである．通常，Fig.2-3 に示す多入力 1 出力の素子でモデル化される．これをユニットと呼ぶ．個々のニューロン間のシナプス結合の強さを結合荷重 W とする．入力 I に対し，結合荷重を乗じたものの総和が，ある閾値 θ を超えると出力関数 f に応じて出力 O を出す．これを数式にすると次式で表される．

$$O = f\{\sum_{k=1}^{n} W_k I_k - \theta\} \tag{2-1}$$

2･2･3　ネットワーク構造

ネットワークの構造は, ネットワークの持つメカニズムの形態から, 大きく二つに分類できる．一つは，層状に結合した階層型ネットワーク，もう一つは，非階層型な相互結合型ネットワークである．相互結合型は Fig.2-5 に示すように各ユニットが相互に結合されているため一度入力された信号がネットワークの内部を何度も伝播するような複雑な構造となる．実際の脳内では，ほぼ相互結合型に近い構造になっていて，各ユニット間の結合荷重は各々適当な時に更新されながら必要な回線だけが強化されていくものと思われる．一方，階層型のニューラルネットワークは，Fig.2-4 に示すように左から右へと信号の流れが決まっている．入力順に計算が行われ，各ユニット間の結合荷重も順次更新される. 入出力関係が分かり易く, 学習アルゴリズムも確立している．音声認識，システム同定，制御などのように，入出力関係が明確な場合には，階層型のネットワークが用いられる．本研究では，時々刻々と変化する対象物の同定および制御にニューラルネットワークを導入する．なるべく複雑な計算を避け，速い制御を行うため，階層型ネットワークを用いることにした．

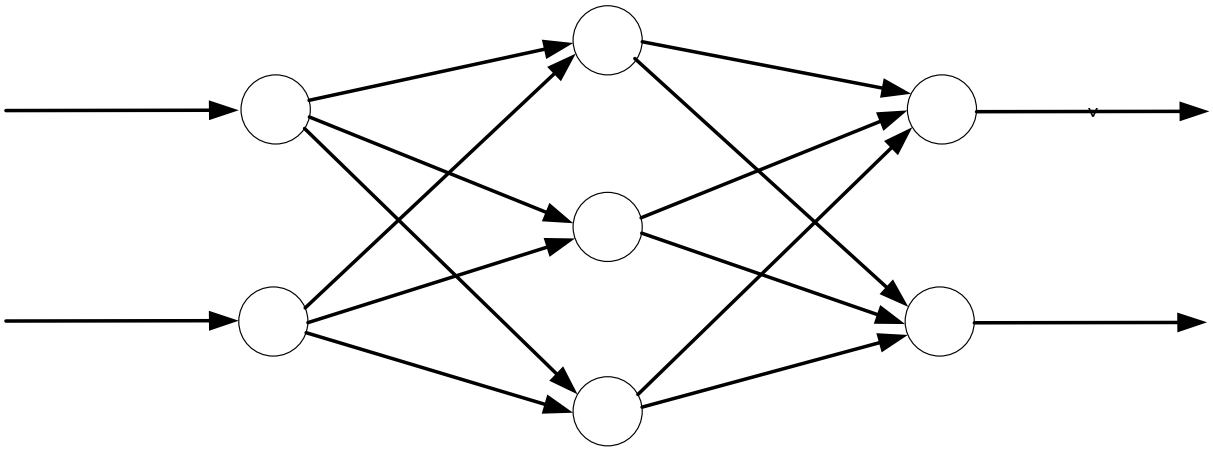


Fig.2-4 Hierarchic neural network

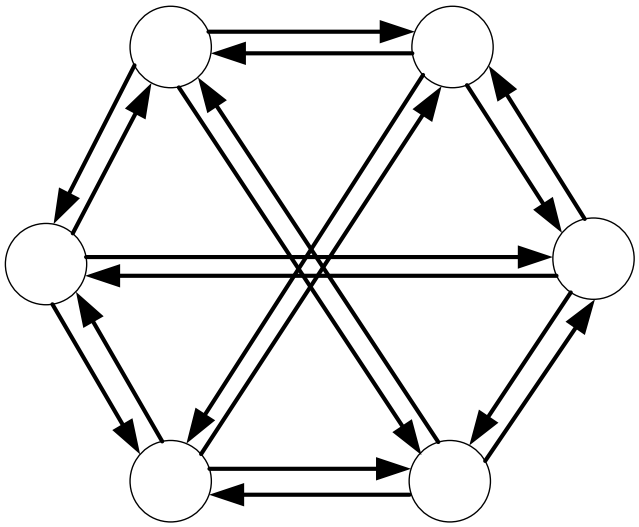


Fig.2-5 Mutual connected neural networks

2･2･4　ニューラルネットワークと学習

階層型のニューラルネットワークが持つ特徴の一つに学習機能が上げられる．

ニューロの学習とは対象とするシステムに合致するようユニット間の結合荷重を調整することである．現在，階層型ニューラルネットワークに最も多く用いられている学習法に誤差逆伝播法（バックプロパゲーション）がある．本研究でもバックプロパゲーションを用いている．そのアルゴリズムには最適化手法である最急降下法を用いられることが多い．入力層，中間層，出力層からなる 3 層のニューラルネットワークのイメージ図を下に示す．

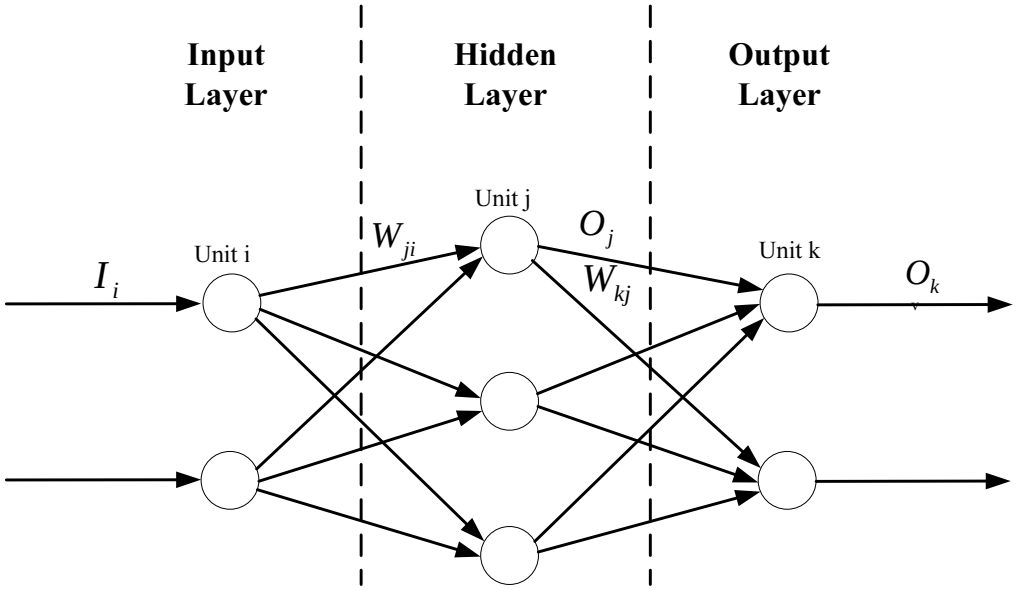


Fig.2-6 Three layered neural networks

$O_j = f(s_j)$	中間層ユニット j の出力
$O_k = f(s_k)$	中間層ユニット k の出力
I_i	入力層ユニット i への入力
W_{ji}	入力層ユニット i と中間層ユニット j 間の結合荷重
W_{kj}	中間層ユニット j と出力層ユニット k 間の結合荷重

2･2･5　シグモイド関数

ニューラルネットワークに用いるシグモイド関数について説明する.
シグモイド関数とは,

$$f(x) = \frac{1}{1+\exp(a-x)} \tag{2-2}$$

と表される関数で$a = 0$の場合のシグモイド関数のグラフを下に示す.

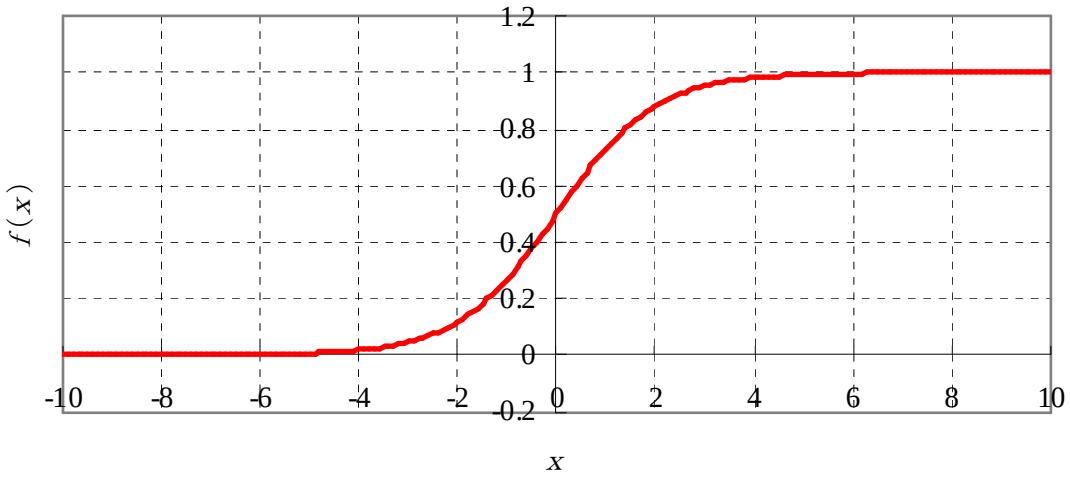


Fig.2-7 Sigmoid function

$$f'(x) = \frac{\exp(a-x)}{(1+\exp(a-x))^2} \tag{2-3}$$

$a = 0$ の場合のシグモイド関数を微分したグラフを下に示す.

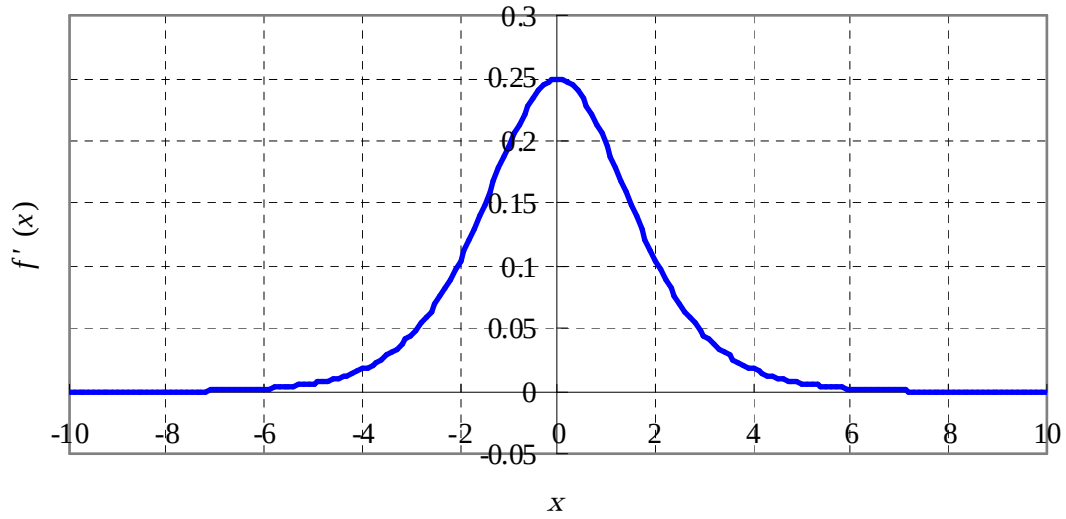


Fig.2-8 Derivative of sigmoid function

$a = 0$ の場合のシグモイド関数とその導関数を並べて表示すると以下のグラフとなる.

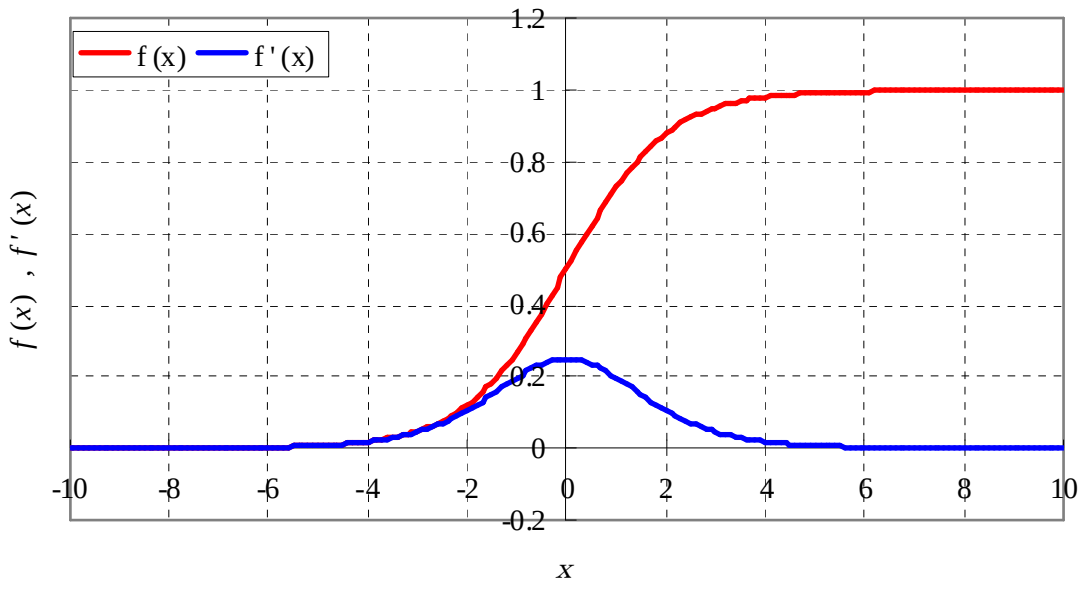


Fig.2-9 Sigmoid function and derivative of sigmoid function

本研究では，フォワードプロパゲーション及びバックプロパゲーションに使用する出力関数には，このシグモイド関数を利用する．

シグモイド関数の特徴としては，

$$f(x) = \frac{1}{1+\exp(a-x)} \tag{2-2}$$

と置くと，上のグラフからも分かるように，

x ≦-5 のとき， f (x)≒0
x ≧5 のとき， f (x)≒1

となり，その出力はほぼ飽和してしまう性質がある．この為，入力データが飽和しないように適切に範囲を修正（スケーリング）する必要がある．また，シグモイド関数からの出力についてもシステムの出力にあわせて，スケーリングする必要がある．

2･3　実験システム

2･3･1　実験装置の構成

本研究で使用する実験装置の構成を Fig.2-10 に示す.

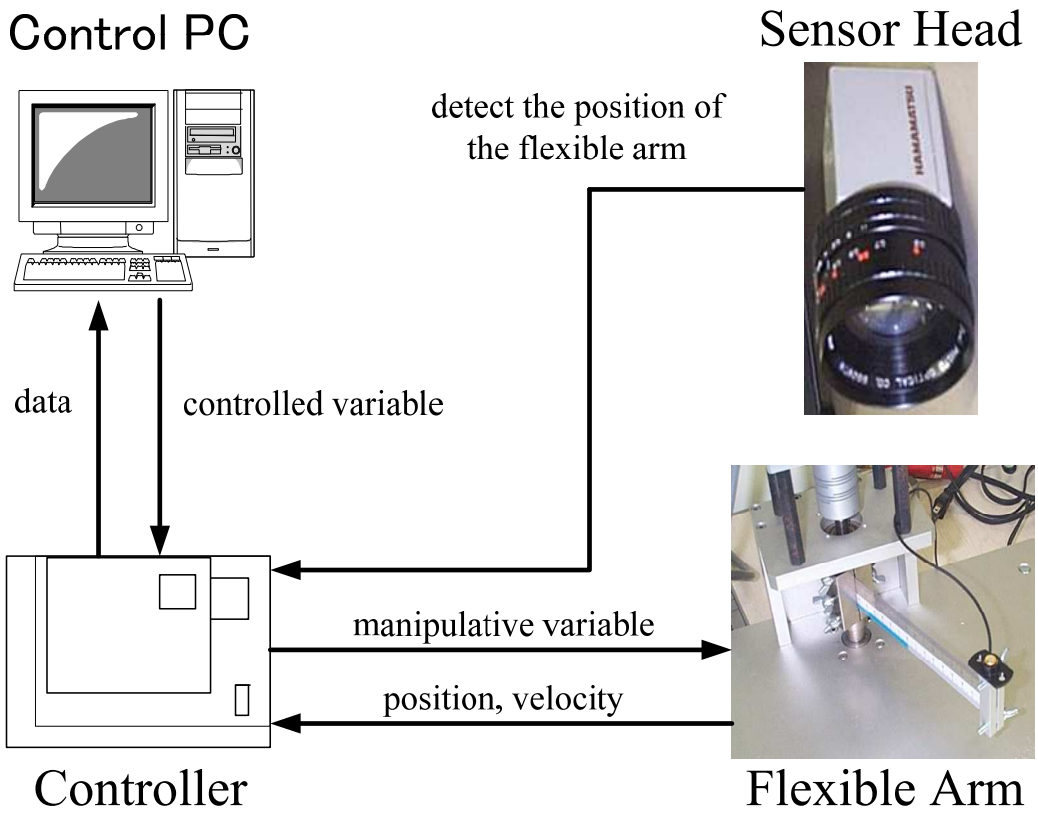


Fig.2-10 Experimental system configuration diagram

2･3･2　実験装置の仕様

本研究で使用する実験装置の仕様を以下に示す.

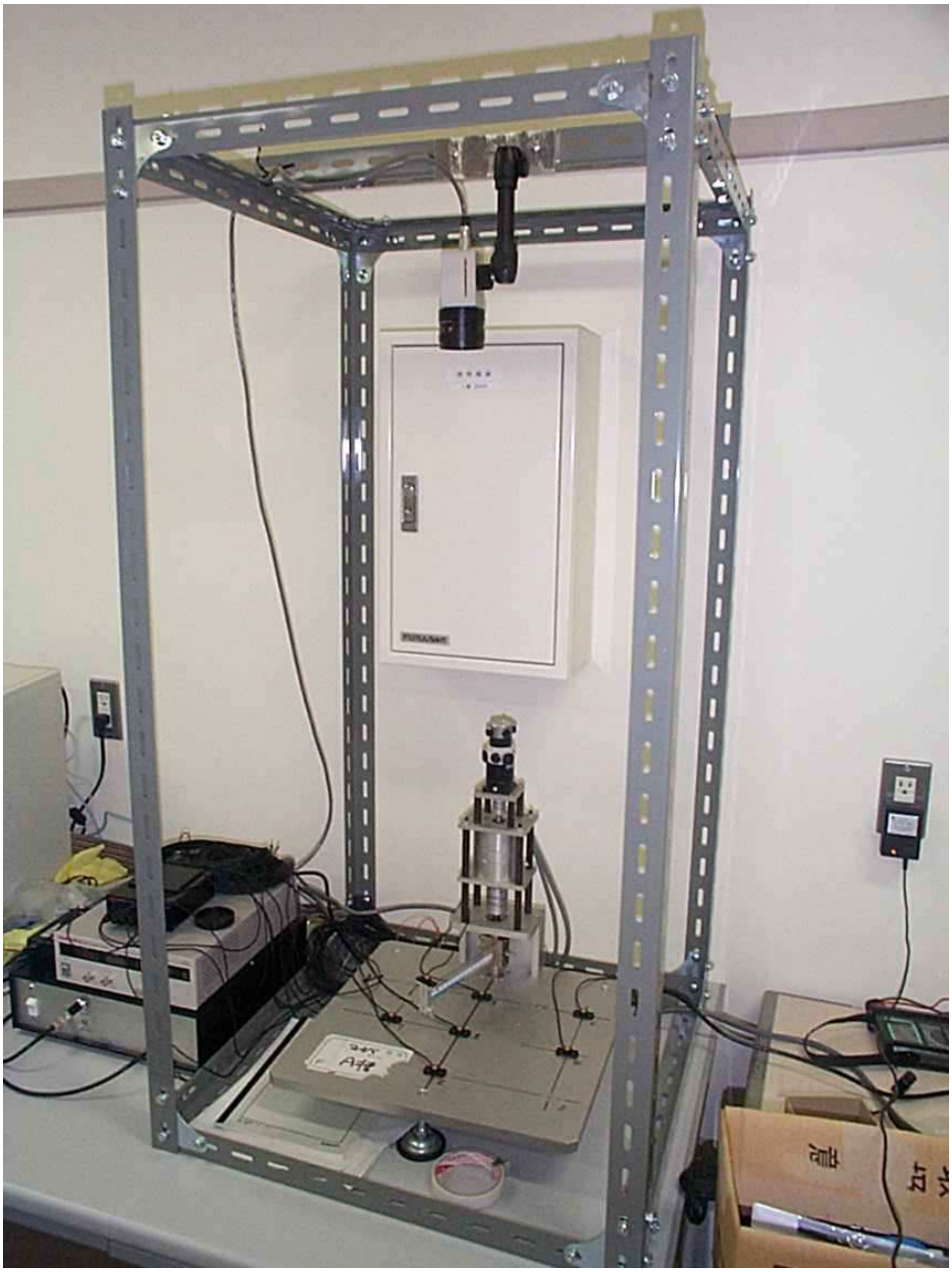


Fig.2-11 Experimental system

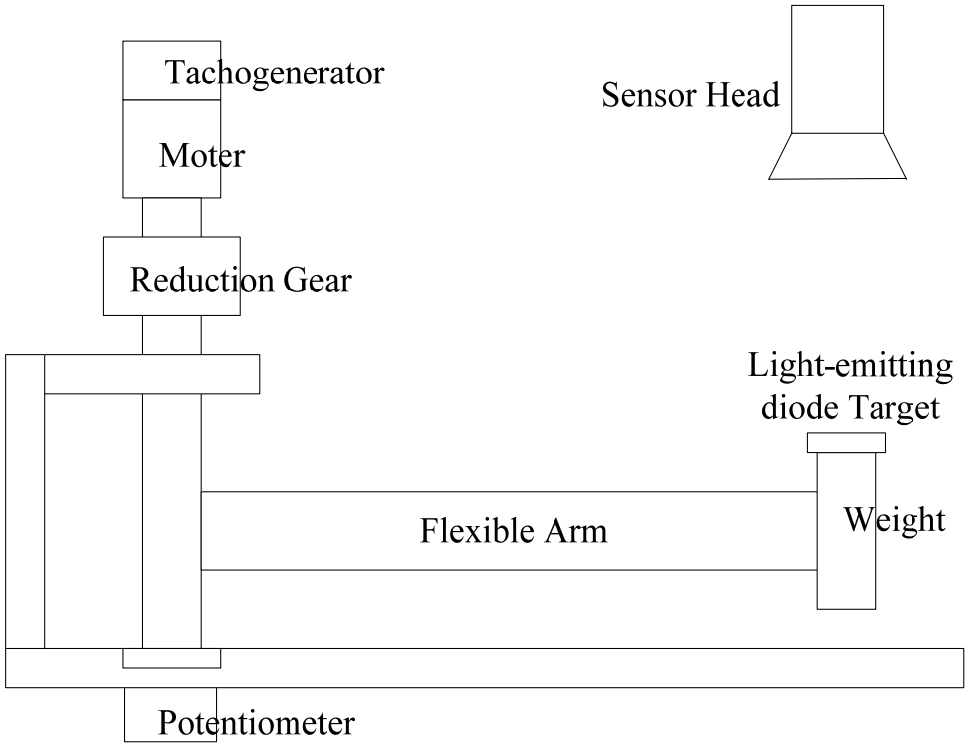


Fig.2-12 System configuration diagram

■サーボモータ（三洋電機製）

モータ，ポテンショメータ，タコジェネレータ，リダクションギヤから構成されている．

■フレキシブルアーム

①塩化ビニール製 18cm 定規（(株）レイメイ藤井製）

②ステンレス製 18cm 定規（(株）YAMAYO）

■ポジションセンサーシステム（浜松ホトニクス（株）製）

フレキシブルアーム先端の重りに取り付けられた LED 発光装置からの赤外線をその上部約 84cm にあるセンサヘッドと呼ばれるカメラにて捕らえ，ポジションセンサーシステムのアンプを通りコンピュータの D／A ボードに位置が電圧として出力される．

１自由度フレキシブルアーム

フレキシブルアームの寸法，構成，材質及び各質量を以下に明記する．

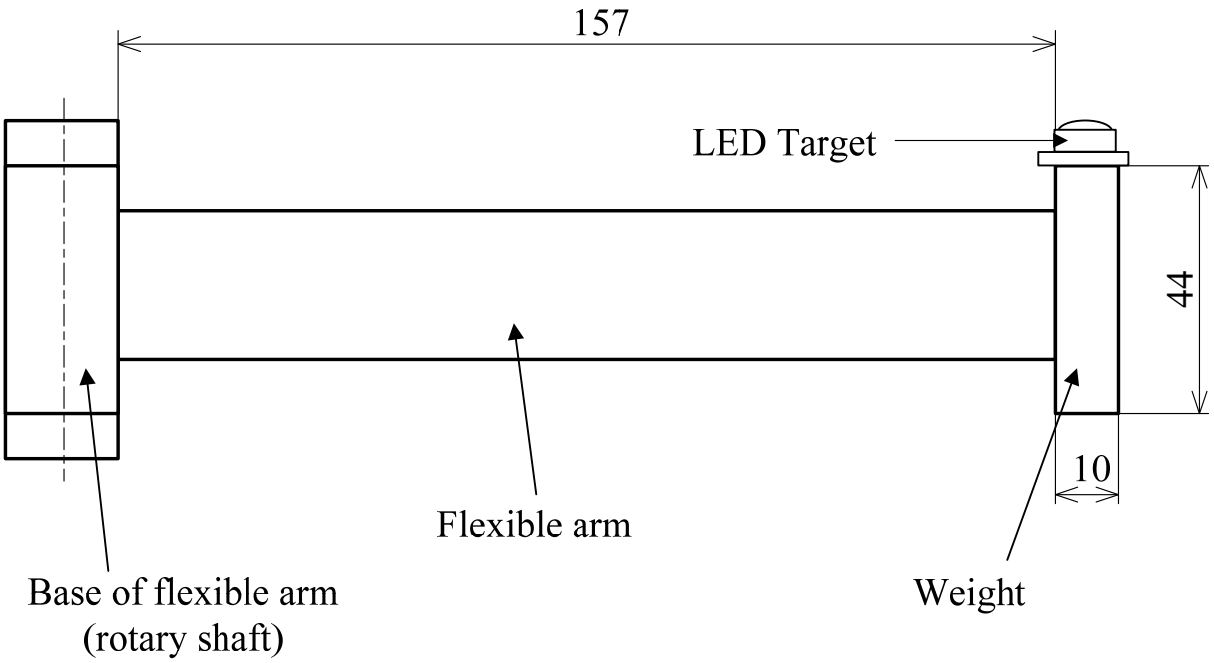


Fig.2-13 Flexible arm –lateral view-

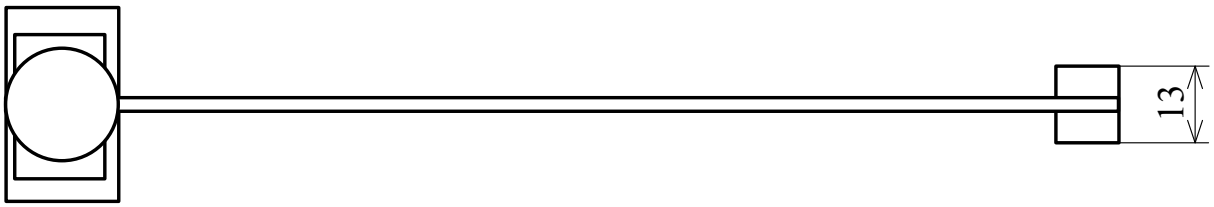


Fig.2-14 Flexible arm –plain view-

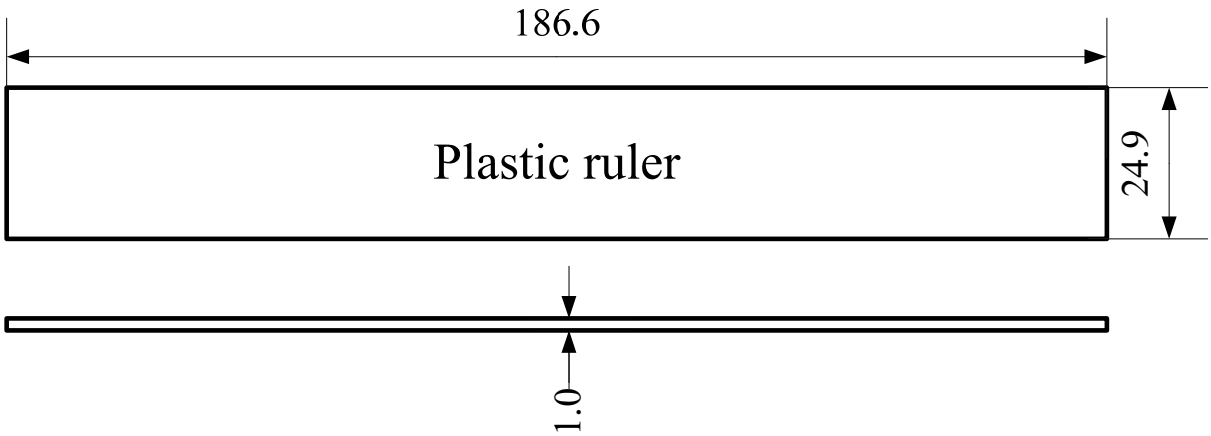


Fig.2-15 Measurement of plastic ruler

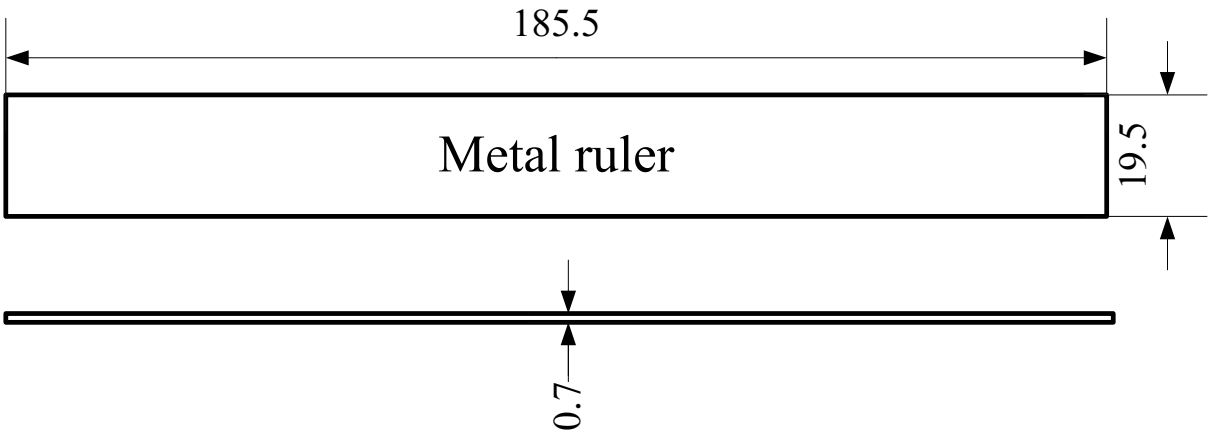


Fig.2-16 Measurement of metal ruler

プラスチック定規
材質　：　塩化ビニール
質量　：　5g

金属定規
材質　：　ステンレス
質量　：　16g

フレキシブルアーム先端の重り
質量　：　25g

位置検出用 LED 発光装置
質量　：　2.5g

モータードライブ

フレキシブルアームをコントロールするために用いるモータードライブの側面写真，構成図，及び各使用を以下に示す.


Fig.2-17 Experimental system

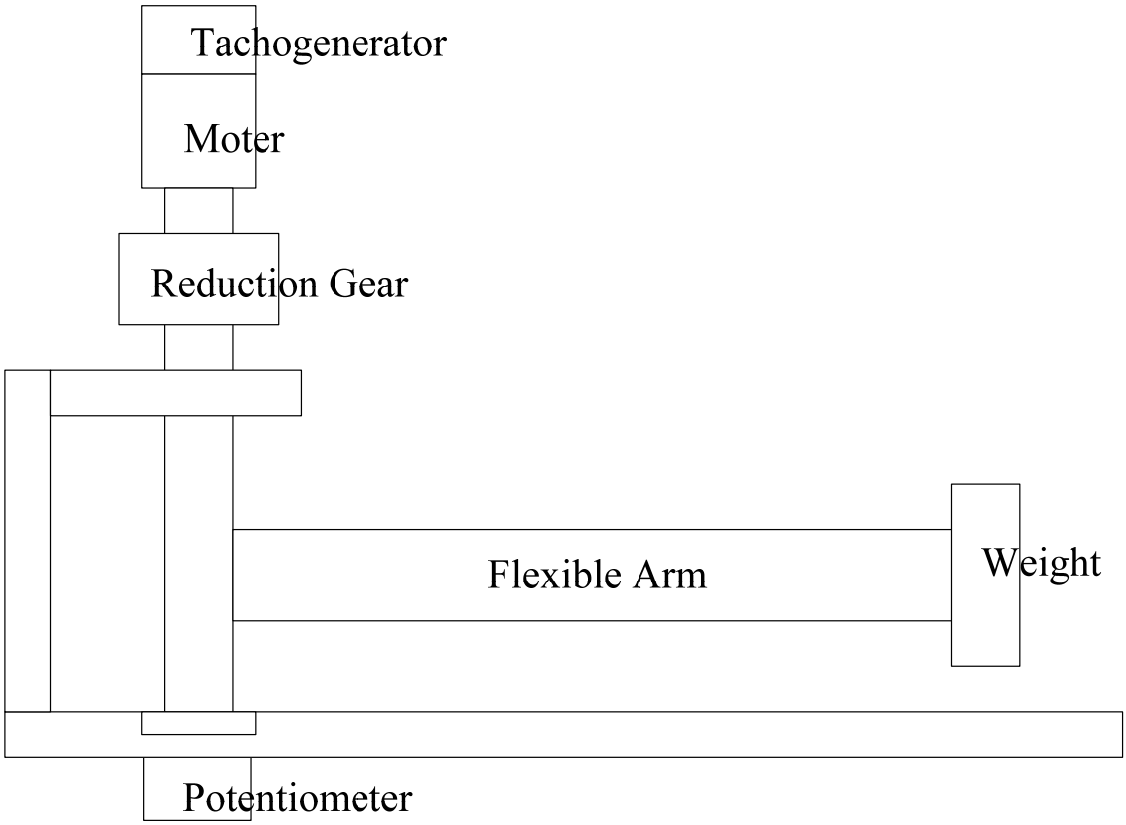


Fig.2-18 System configuration diagram

主要定格及び仕様

形式	μSWA-2	備考
定格出力電圧　±V	80	電源 AC100V の時
定格出力電源　±A	2.2	
最大出力電圧　±V	88	電源 AC100V の時
最大出力電源　±A	5.5	
入力電源	AC20～110V,　DC30～150V	
主回路	パワーMOSFET,PWM(25KHz),可逆	
出力回路	LC フィルター内蔵(負荷短絡抑制機能　1 分間)	外部リアクトル不要
絶縁耐圧	主回路，信号間 1200V1 分間	
減定格	95%以上	
動作温度，湿度	0~50℃，85%RH 以下(結露なし)	
保存温度，湿度	-20～85℃，85%RH 以下(結露なし)	
外形寸法，重量	120W x 210L x 42D　1.2kg	

制御部仕様

モータドライバユニットは，指令電圧を受け取って，モータに対して駆動するための電力を与え，サーボモータのタコジェネレータからデータを受け取ってそれを I/O 形式で出力する．モータドライバユニットの仕様を以下に示す.

項目		定格及び仕様	備考
指令入力		0～±10V	
指令インピーダンス		100kΩ	
速度帰還電圧		±2~50V(標準出荷 7 V/Krpm,3000rpm)	位置制御帰還電圧 ±10V　MAX
電流応答		500μS 以内(10~90%ステップ応答)	
速度制御範囲		5000:1 以上(タコゼゼネレータ 7V/Krpm の時)	
負荷変動		±0.1%以内	10~100%
分解能		速度制御系，位置制御系，0.02%以下 電流制御系，1%以下	
直線性		電流制御系　３％以下	
設定電源(内蔵)	安定度	±0.25%	
	出力電圧	±12V/±0.5V(出力短絡防止 510Ω内蔵)	出力オープンにて
内蔵機能	入力信号	全停止，正転禁止，逆転禁止，ゲイン低下，リセット	
出力信号	アラーム（ヒートシンク加熱 75℃以上にて出力）		オープンコレクタ出力
保護機能	電流制限，電圧制限，ヒートシンク加熱遮断		
表示ランプ	POWER(“POW”LED 表示),ALARM(“ALM”LED 表示)		
オフセット	速度ゼロ（速度指令 0 の時）[OFFSET]		VR0
電流制限	5～100%　[laFs]		VR2
速度	速度フルスケール　[SPEED]		VR
比例要素	PI 制御の”P”調整　[P]		VR3
積分要素	PI 制御の”I”調整　[DSI]		
電圧制限	0～100%　　[Va]		VR5
ゲイン	0～20 倍　　[GAIN]		VR4
電流オフセット	電流 AMP ゼロオフセット調整　[lo]		調整不要
電流ゲイン	電流 AMP　MAX 調整　[lg]		調整不要

ポジションセンサシステム

フレキシブルアームの先端に取り付けられた LED の発光装置が赤外線を発し，その上部からカメラにて赤外線を検出する．これによって得られる位置データはポジションセンサアンプで増幅され，制御用コンピュータへ送られる．以下にポジションセンサとアンプの使用を示す.

LED ドライバ仕様

LED ドライブ点数	最大 7 点
LED 点灯周波数	300Hz　（146~300Hz 設定可能）
使用電源	内部：単三電池　　／1.5V x 4 本 外部：DC4.5V～7V
連続使用時間	約 20 分（LED7 点接続時） （Ni-Cd 電池使用時）
消費電流	1A（LED7 点接続時）

ポジションセンサ仕様

一般仕様

検出器	半導体位置検出素子（S1880）
使用受光面寸法	10mm x 10mm
レンズマウント	C マウント
動作周囲温度	0℃～+40℃
保存周囲温度	-10℃～+50℃
動作，保存周囲湿度	90％以下（結露しないこと）
入力電源	100V
外形寸法及び重量センサヘッド	40(W) x 42(H) x 64(D)mm　　約 140 g
コントローラ	232(W) x 74(H) x 308(D)mm　約 3.4 kg

電気的仕様

出力電圧：X 軸 Y 軸	-5V～+5V -５V～+５V
出力インピーダンス	500Ω±50Ω
外部クロック信号	TTL レベル
サンプリング周波数	内部モード：300Hz（標準）
推奨測定光量	光量レベル（Σ）= 4～8
位置検出誤差：ZONE　A ZONE　B	±1% ±2%
光量変化による誤差	±1%（Σ8→4）
分解能	1／5000
ジッタ	±1／1000
ドリフト	±0.5%／DAY（ただし初期 30 分の変動を除く）

制御用ボード

a. A/D ボード
カメラセンサアンプからのアナログ信号をデジタル信号にしてパソコンに送る
（株）コンテック　　AD-12-16(PCI)
　使用時の入力レンジ　：　±5V
　分解能　　：　12bit

b. D/A ボード
パソコンからのデジタル信号をアナログ信号にしモータドライバに，モータへの指令値を送る．
（株）コンテック　　DA-12-16(PCI)
　使用時の出力レンジ　：　±5V
　分解能　　：　12bit

c. PIO ボード
モータドライバよりデジタル信号の送受信を行う．
（株）コンテック　　PIO-48D(PCI)

制御用パソコン

DELL Dimension XPS T800r
Intel 社製 PentiumⅢ 800MHz 搭載

2･4 まとめ

本章では，ニューラルネットワークに関する概要と実験装置の構成，仕様について説明した．

第3章
学習結果Ⅰ　～人間と機械の協調学習～

3･1　概説

本章では，はじめに学習制御を用いない一般的な PID 制御を用いた場合のフレキシブルアームの振動制御について述べる．その後，ニューラルネットワークを用いた学習制御を適用する．機械が学習する過程では様々な形で人間が教示できることを述べると共に，機械学習における人間と機械の望ましい関係を示す．

3･2　実験内容

1. フレキシブルアームの根元に強制振動を加える．
2. 外力を外し自由振動となった後，振動制御を行う．

Fig.3-1 は実験全体の流れをグラフで示したものである．

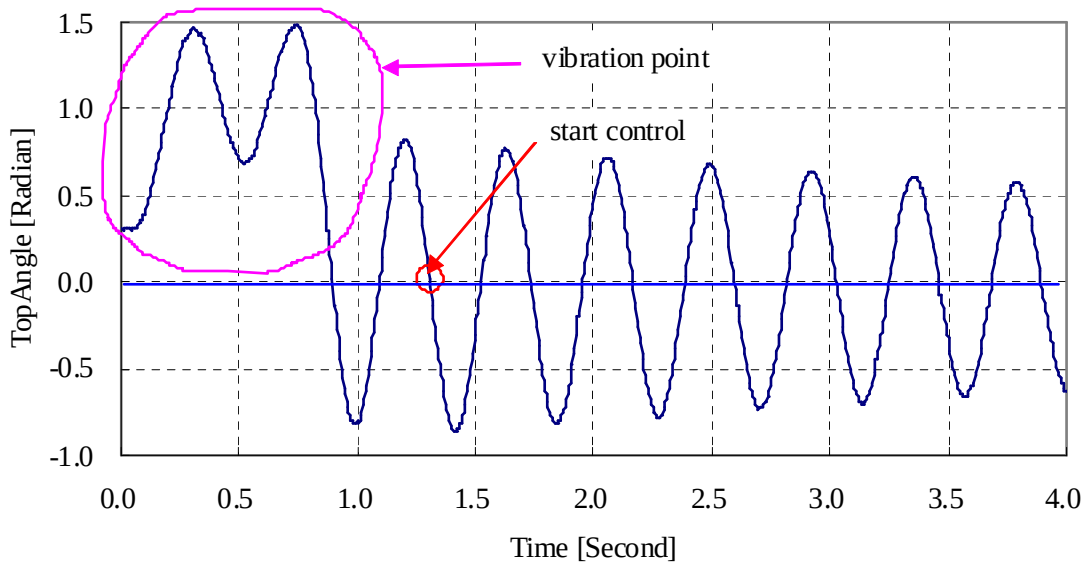


Fig.3-1 End motion of flexible arm

0.000 ⇒ 0.700 [sec]
フレキシブルアームの根元にあるサーボモータに強制振動を加えるための指令を与える．
0.700 ⇒ 1.305 [sec]　自由振動をさせる．
1.305 ⇒ 30.000 [sec]　振動制御を行う．

3･3　基本制御（非学習制御）

本節では一般的な PID 制御により振動制御を行った結果と考察を述べる．

3･3･1　P 制御（Proportional Control）

フレキシブルアームは柔軟媒体である為，外力が加えられると歪エネルギーが発生する．そこで，フレキシブルアームの根元の角度$\hat{\theta}(k)$と先端の角度$\theta(k)$の情報を利用し，この誤差$e_p(k)$が小さくなるようなフィードバック系を組む．こうすることで，フレキシブルアームの撓みを小さくするような制御入力が加わり，歪エネルギーを減少させることによって制振させることを考えた．Fig.3-2 にフレキシブルアームの振動時のイメージを，Fig.3-3 にブロック線図を示す．

目標角度　：　$\theta(k)$

誤差　：　$e_p(k) = \theta(k) - \hat{\theta}(k)$

制御入力　：　$u(k) = Kp \times e_p(k)$　（Kp　は比例ゲイン）

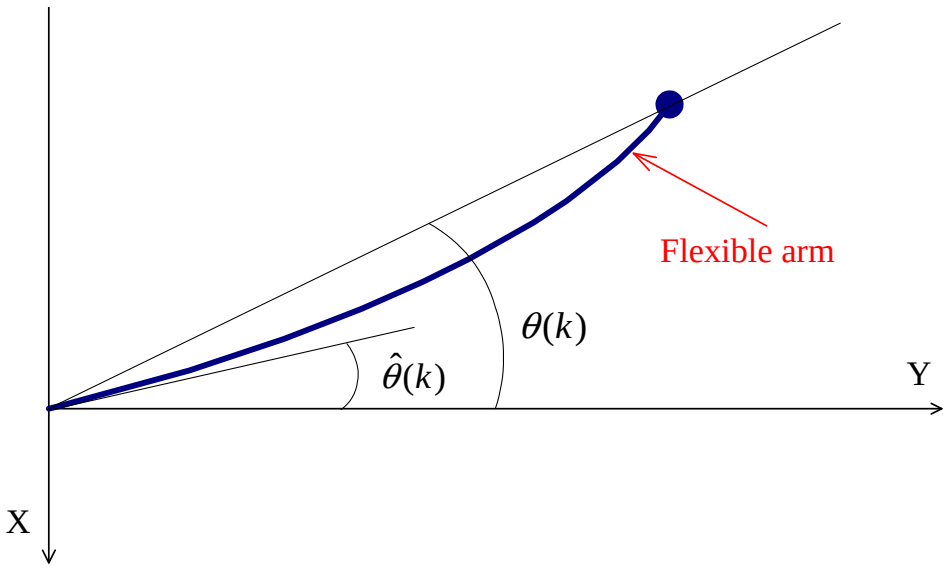


Fig.3-2 Flexible arm diagram

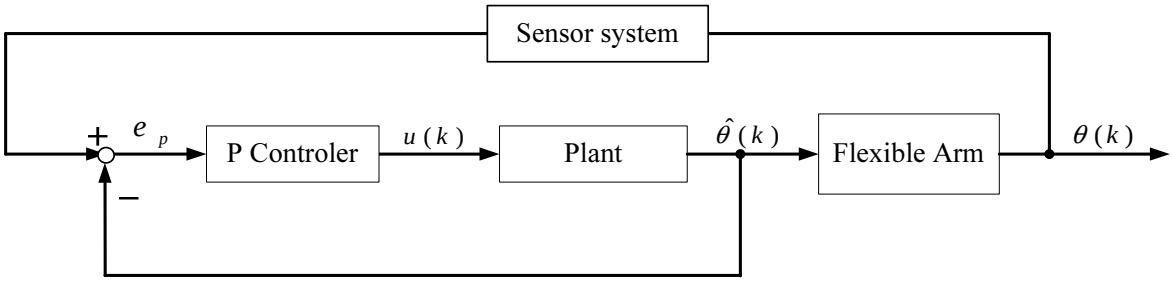
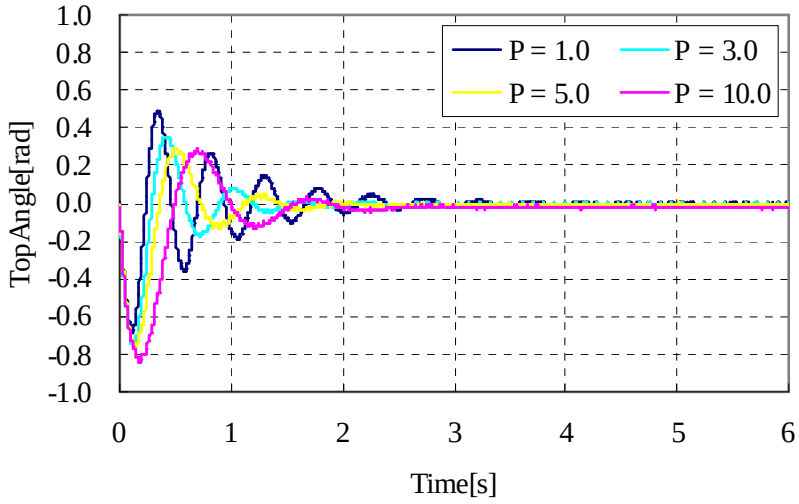
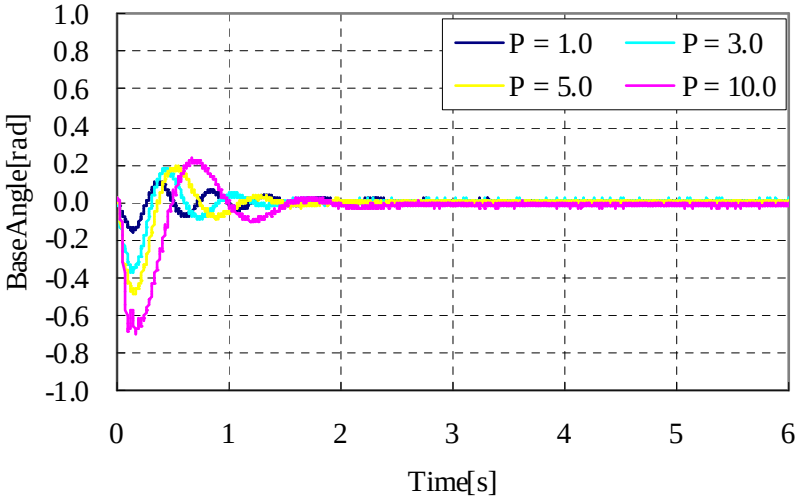


Fig.3-3 Proportional control block diagram

以下に，比例ゲインを変化させた時の，先端角 $\theta(k)$ 及び根元角 $\hat{\theta}(k)$ の様子を示す．

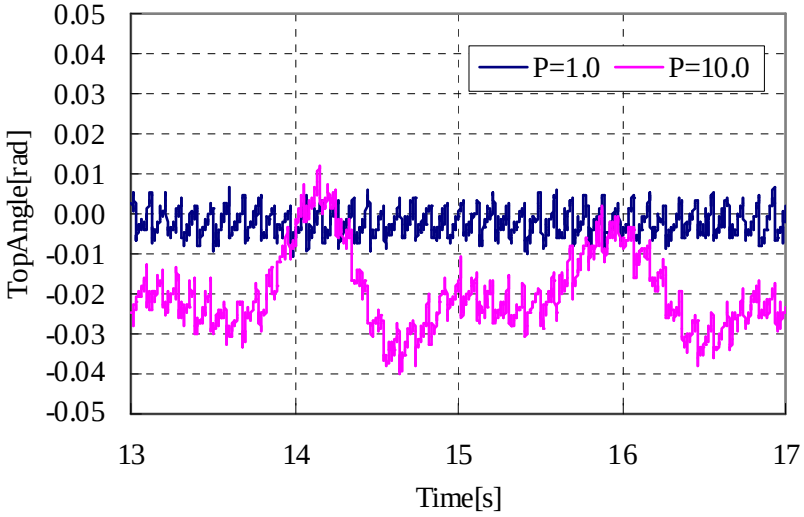


(a) Top angle $\theta(k)$

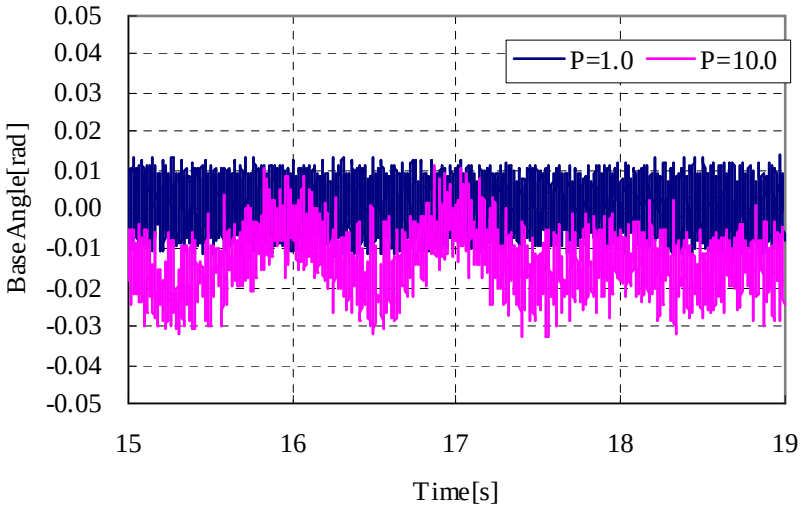


(b) Base angle $\hat{\theta}(k)$

Fig.3-4 Experimental result -overall behavior-



(a) Top angle $\theta(k)$



(b) Base angle $\hat{\theta}(k)$

Fig.3-5 Experimental result - stationary behavior-

Fig.3-4 を見ると，比例制御により先端角，根元角共に，一定値に収束していることがわかる．また比例ゲインを大きくするほど収束性能が向上しているのがわかる．一方 Fig.3-5 には定常時の先端角，根元角の挙動を拡大して示すが，比例ゲインが大きいほど定常偏差が残り，かつ先端角，根元角が微振動している様子(残留振動)がわかる．

ここでは，比例ゲインの設定において「全体の大きな挙動を抑える」ということと，「定常時の微振動を抑える」ということがトレードオフ関係になっていることがわかる．

3･3･2　PD 制御（Proportional-Derivative Control）

位置制御の項

目標角度　：　$\theta(k)$

誤差　：　$e_p(k) = \theta(k) - \hat{\theta}(k)$

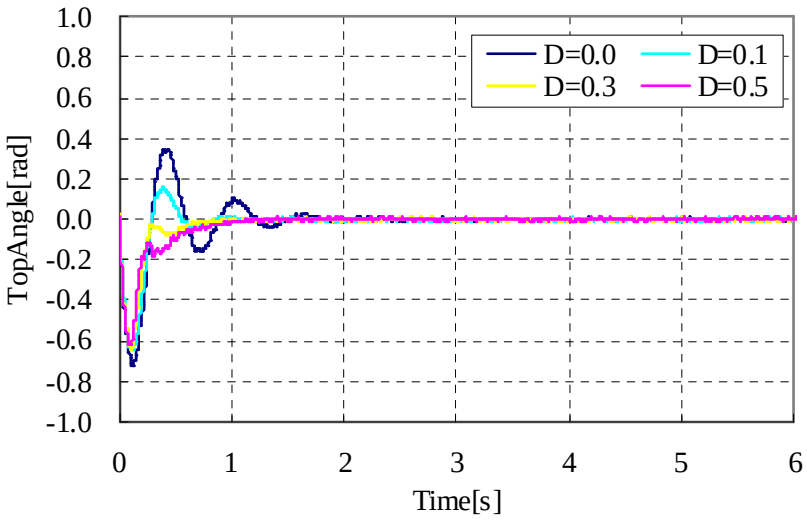
速度制御の項

目標角度　：　$\theta_d = 0$

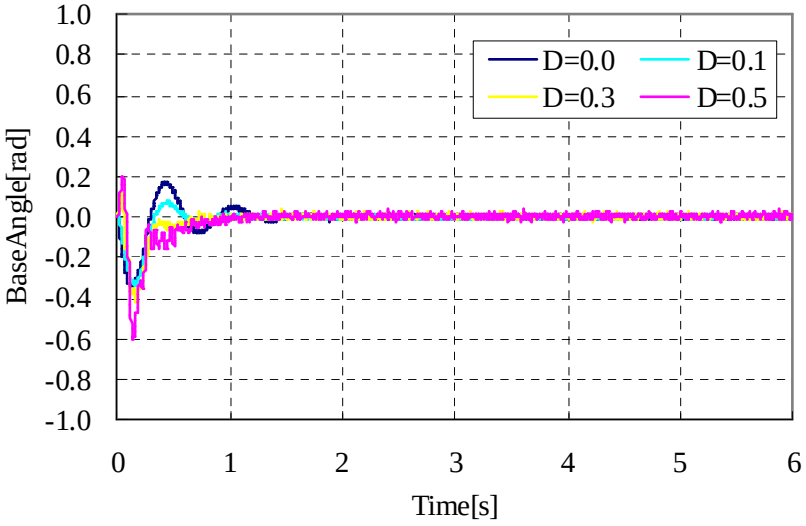
誤差　：　$e(k) = \theta_d - \theta(k)$

位置制御　＋　速度制御　による制御入力

制御入力　：　$u(k) = Kp \times e_p(k) + Kd \times \dfrac{e(k) - e(k-1)}{T}$

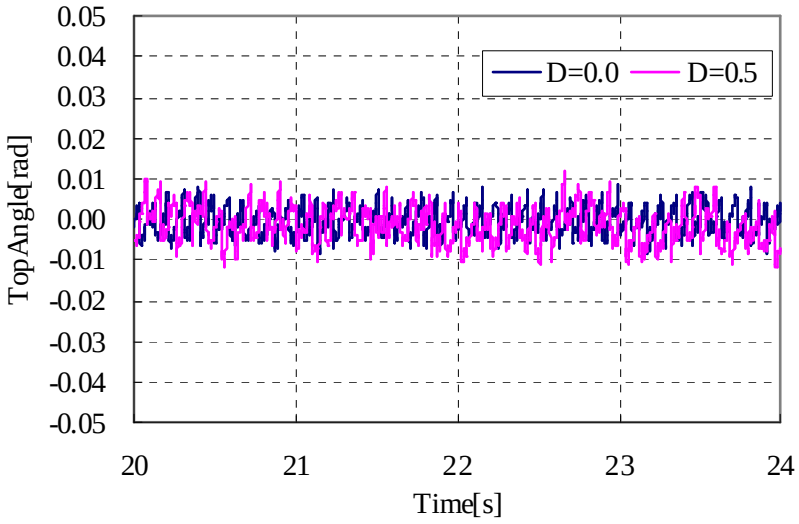


(a) Top angle　$\theta(k)$

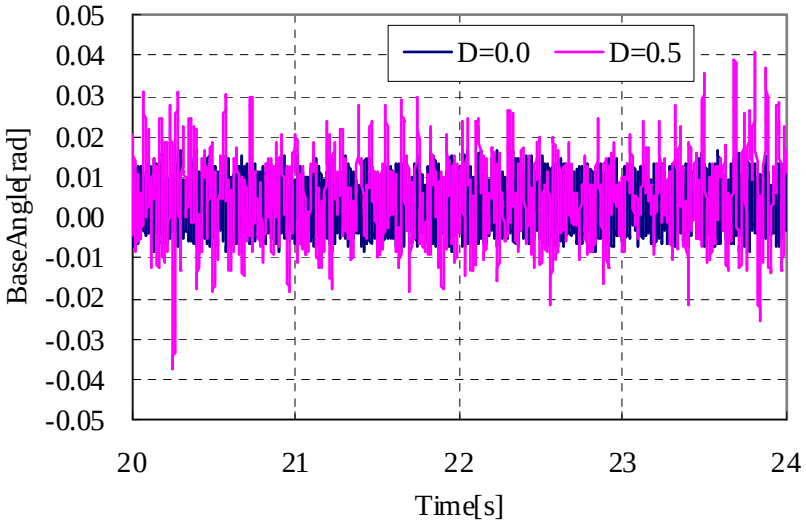


(b) Base angle　$\hat{\theta}(k)$

Fig.3-6 Experimental result -overall behavior- (P=3.0)



(a) Top angle $\theta(k)$



(b) Base angle $\hat{\theta}(k)$

Fig.3-7 Experimental result - stationary behavior- (P=3.0)

Fig.3-6 および Fig.3-7 は微分制御を加えたことによる，先端角，根元角の動きの比較結果である．制御入力として微分項を加えることで速度エネルギーを素早く取り除き，収束を早めていることがわかる．またオーバーシュートをおさえていることもわかる．一方，定常時の残留振動は微分ゲインが高いほど振幅が大きくなっており，前述のトレードオフ問題がここでも生じていることがわかる．

3･3･3　PID 制御（Proportional-Integral-Derivative Control）

位置制御の項

目標角度　：　$\theta(k)$

誤差　：　$e_p(k) = \theta(k) - \hat{\theta}(k)$

速度制御の項

目標角度　：　$\theta_d = 0$

誤差　：　$e(k) = \theta_d - \theta(k)$

積分制御の項

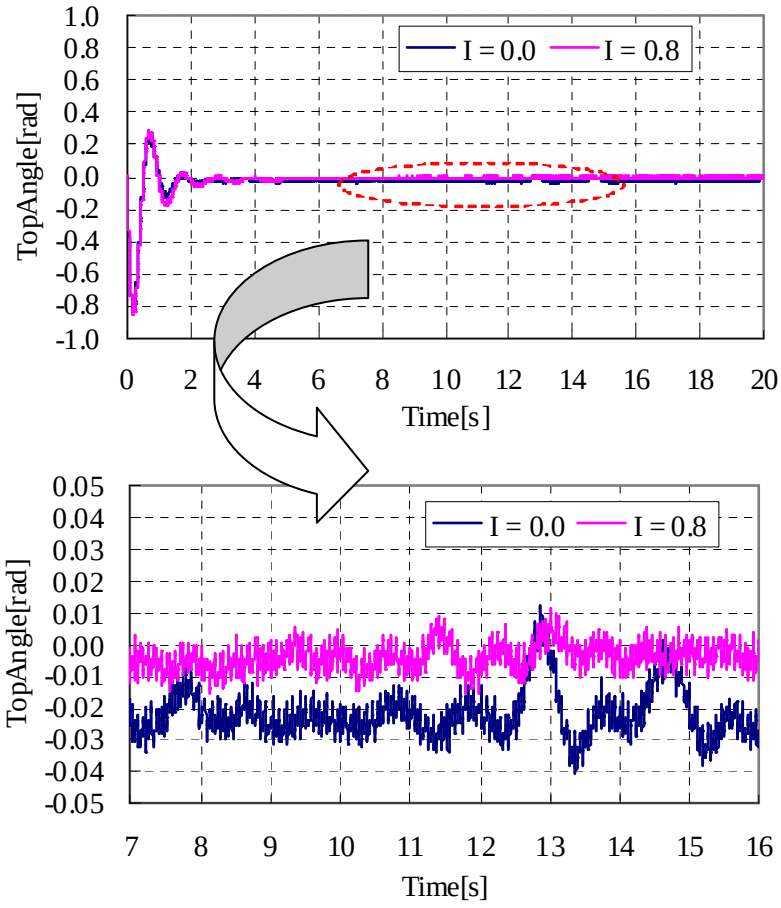
目標角度　：　$\theta_d = 0$

誤差　：　$e(k) = \theta_d - \theta(k)$

位置制御　＋　積分制御　＋　速度制御　による制御入力

制御入力　：　$u(k) = Kp \times e_p(k) + Ki \times \sum_{j=-\infty}^{k} e(k)T + Kd \times \frac{e(k) - e(k-1)}{T}$

Fig.3-7 を見ると，積分項を加えたことで定常偏差が小さくなっていることがわかる．



Top angle　$\theta(k)$

Fig.3-8 Experimental result - stationary behavior- (P=3.0 D=0.8)

3･4　学習制御

3･4･1　問題設定

前節までで，一般的な制御手法である P 制御，PD 制御，PID 制御によりフレキシブルアームの振動を抑制することが可能であることを示したが，以下の 2 つの問題点を有する．

①比例，微分，積分ゲインの調整が膨大なトライアンドエラーにより決定している
②全体の大きな振動の収束を速める為にハイゲインにすると，定常時の残留振動が発生する

この問題に対して，①は**ゲインの調整を効率的に自動で行いたい**と考えるのが自然である．また②については，全体の大きな振動抑制と，定常時の残留振動の双方を小さくする**トレードオフ問題を自動的に調整したい**という考えに至る．

そこで，機械学習の問題設定として，これら①②の 2 つの課題をフレキシブルアームロボットが自動的に調整する制御器（コントローラ）を学習することを考える．その過程で，人間がどのような形でロボットの学習に関わっていくのかを述べる．

3･4･2　人間と機械の接点：教示

本タスクにおいてフレキシブルアームロボットが，振動を抑制するための最適な制御を学習するにはどのような情報が必要であろうか？ ロボットにとって知るべきことは 2 つある．

一つは，どのような制御入力（モータ指令電圧）を与えると，どのような出力（先端角，根元角）が得られるか？という，入出力の関係を理解する**＜モデリング＞**の問題である．もう一つは，そのモデルに対して，時々刻々どのような入力を行えば所望の制御が実現できるかという**＜コントロール＞**の問題である．（両者の関係から，コントロールはモデリングの逆問題とも言える．）

ここでいう所望の制御とは，評価関数の事を意味しており，人間が与えなければならない．例えば「二乗誤差が最小となるような制御」や「整定時間が最小となる制御」などである．この評価関数自体は機械自ら学習することはできない，という点は序章で述べたとおりである．

一方，与えられた評価関数の基，モデリング，コントロールを機械が学習する過程では，人間と機械の間にインタラクションの融通性が生じる．すなわち人間は機械の学習に対して介入することができ，介入の度合いを任意に決定することができる．例えばモデリングでは，制御対象の構造（次数や線形・非線形性）などの情報を詳細に機械に教えた状態で学習させてもよいし，前提を全く与えなくても学習させること自体は可能である．コントローラに関しても同様である．

一般に，機械へ与える教示量を多くすればするほど，機械学習としての自律性は損なわれ，学習の結果出来上がったコントローラは人間の想像の域を超えないことが多い．逆に教示量を少なくすればするほど，機械の自律性は大きくなるが，学習時間が長くなり，場合によっては所望の制御が実現できないことがある．そこで次節以降では，フレキシブルアームロボットが振動抑制を行うコントローラを学習するという問題設定に対して，人間がどのような情報を与えることができるか，また，人間と機械はどのように関わっていくことが望ましいのか？について詳細を述べる．

3･5　学習の為のモデリング

3･5･1　モデリング概要

3･3 節で挙げた①②の 2 つの課題を解決するコントローラの学習を行うためには，制御入力に対してどのような出力が返されるかを模擬する制御対象のモデル（シミュレータ）が必要となることを述べた．モデリングには大きく，物理原則に基づく第一原理モデリング，実験入出力データから内部ブラックボックスの推定を行うシステム同定モデリング，両者の中間に位置する，既知の式構造のみを与えて係数を推定するグレイボックスモデリングなどがある．自律的な学習という意味においては機械自ら何らかの形で制御空間のモデルを獲得し，そのモデルに対してコントローラを設計することが期待される．ここではモデル獲得の際の人間と機械とのインタラクションを議論するため，モデル自由度が高いシステム同定モデルをベースにして学習を進める．

3･5･2　システム同定

システム同定とは，対象とする動的システムの入出力データの測定値から，そのシステムと入力に対する出力が同等であるモデルを作成することを意味する．詳細は文献[220,221]を参照されたい．

以下にシステム同定のエッセンスのみ記載する．

Step1. 同定実験の設計

･同定入力，サンプリング周期などの選定

Step2. 同定実験

･同定対象の入出力データの収集

Step3. モデル構造の選定

･線形，非線形　･連続，離散

･パラメトリックモデル，ノンパラメトリックモデル

･モデル次数の決定

Step4. システム同定

･入出力データをモデルに当てはめる

Step5. モデルの妥当性の評価

･時間領域での評価

･周波数領域での評価

本研究における入力と出力は以下のとおりである．

入力：モータ指令電圧 *u(k)*

出力：根元角 $\hat{\theta}(k)$ (モータ部分) 先端角 $\theta(k)$ (フレキシブルアーム部分)

システム同定を行う上で一番大切な作業は入力データの選定である．入力データの検討に際しては，どの程度の正弦波成分を含ませる必要があるかという周波数領域の検討と，振幅をどの程

度にしたらよいかという時間領域の検討が含まれる．ここで適切な入力データを与えないと，機械がコントローラを学習する為に必要な有益なモデルが作成できない．つまり**入力データ**の選定は，人間が機械に与える重要な**教示**の一つと言える．これは機械に対して，学習の為の教科書を渡すようなものであり，英語の試験を受ける生徒に国語の教科書を渡しても望ましい学習はできないのと同じである．

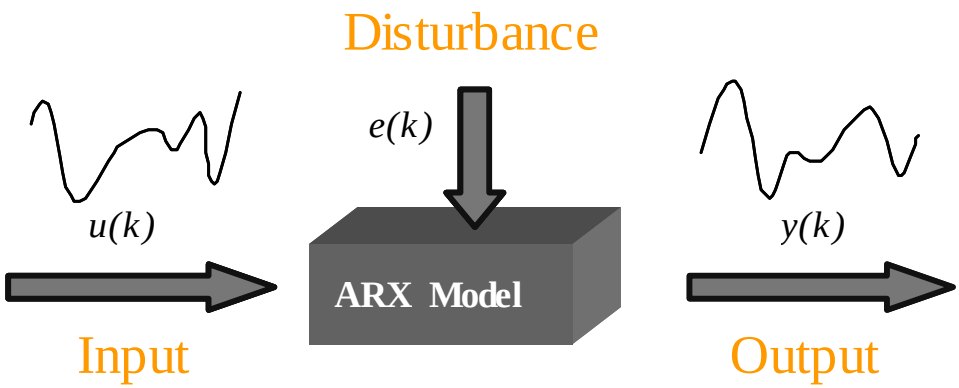
次にサンプリング周期の選定についてだが，これまで同定に最適なサンプリング周期に関する研究がなされており，サンプリング周期によりモデルの良し悪しが大きく変化することが知られている．この為，**サンプリング周期**の選定も，人間が機械に与える**教示**の一つである．

同定入力データ，サンプリング周期の検討が終わったら，実システムに入力を与え出力データを収集する．この入出力データから機械がどのような構造のモデルを作り出すか？という**モデル構造**の選定も**教示**となる．このように機械が制御空間のモデル一つ学習するにしても，人間が機械学習過程に介在する余地が非常に多いことがわかる．

3･5･3　ARX モデル

システム同定モデルにおいて最も基本的なモデルにARXモデルがある．ARXモデルとはFig.3-9に示すように，「システムの出力は過去の入力と出力からなり，それらの線形和であらわすことができる」ということを数式にしたものである．システム同定では実験から得られた入出力データをもとに係数 $a1,a2\cdots,b1,b2\cdots$ を推定していく．ここで**モデルの次数 *n***も**教示**の一つとなる．制御対象のモデル次数を把握することは，制御系設計においては安定性や整定性を評価する上で非常に重要な情報であるが，機械学習における次数の選定は，機械が学習モデルを限定するための重要な教示情報になる．

Fig.3-10 は本制御対象のモータ部分のボード線図を実験的に描いたものである．ボード線図からわかることは，対象の次数が 1 次あるいは 2 次であることが見て取れる．また，-3[dB]の時の角周波数は 2.0[Hz]であるが，この角周波数を超える値でモータ部分を使用することは制御上，追従性が悪化する可能性があることを意味している．（制御に有用な周波数帯を帯域幅と言う．）



$$y(k) = a_1 * y(k-1) + a_2 * y(k-2) + \cdots + a_n * y(k-n)$$
$$+ b_1 * u(k-1) + b_2 * u(k-2) + \cdots + a_n * y(k-n) + e(k)$$

Fig.3-9 ARX Model

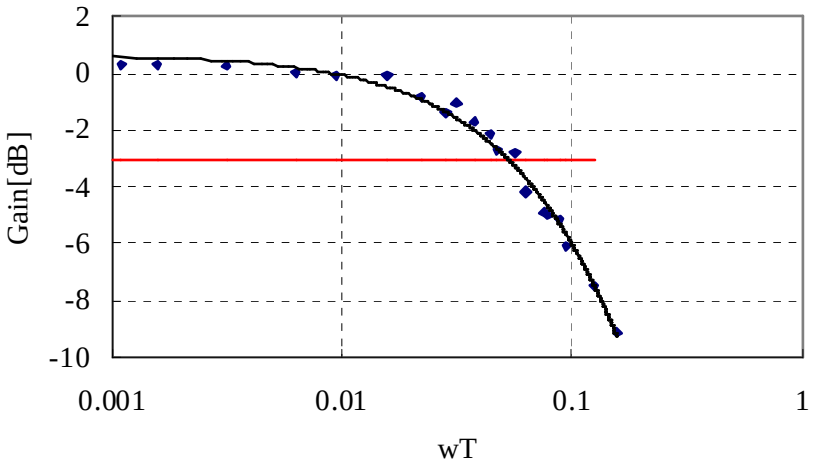


Fig.3-10 Bode Block Diagram –motor-

この**帯域幅**の情報も重要な**教示**である．帯域幅をロボットに教示するということは，ロボットは制御上問題となり得る帯域を外して，限定された空間の中で最適なコントローラを探索することができるということである．これは人間の学習で言えば，予備校の教師が生徒に試験に出る重要ポイントだけを教えて，効率的に学習させることと近い．

以上述べたように，モデルへの前提知識の与え方に人間と機械のインタラクションの自由度がある．機械学習における人間と機械の望ましい関係を議論することは，これらの自由度のある教示に対して，どのような教示を与えることが望ましいのか？を議論することに他ならない．

3･6　教示と汎化性

次に機械学習における重要な問題である「教示」と「汎化性」の関係について述べる．

先に述べたように，モデルの帯域幅を教示することは，ロボットが学習する探索空間を狭めるという意味で学習を効率化させる要因であることを述べた．しかしこのことは裏を返すと，探索空間以外の周波数領域では，ロボットは適応できなくなる可能性があることを意味している．これを以下に具体的に示す．

Fig.3-10 から本システムの帯域幅は 0.0～2.0Hz 程度であることがわかっている．ここではそれらを包括する 0.0～5.0Hz の周波数を満遍なく含むバンドノイズを入力データとしてモデルを作ることを考える．Fig.3-11 に時系列データを，Fig.3-12 に周波数データを示す．この入出力データを用いて二次の ARX モデルを作成した. モデリング結果を Fig.3-13 に示す. Fig.3-13 に示すように，ARX モデルにより十分モデル化ができていることがわかる．モデル式は下記のとおりである．

$$\hat{\theta}(k) = 0.6936\hat{\theta}(k-1) + 0.2385\hat{\theta}(k-2) + 0.0462u(k-1) - 0.0187u(k-2)$$

（なお，同定には MATHWORKS 社の MATLAB System Identification Toolbox を用いた.）

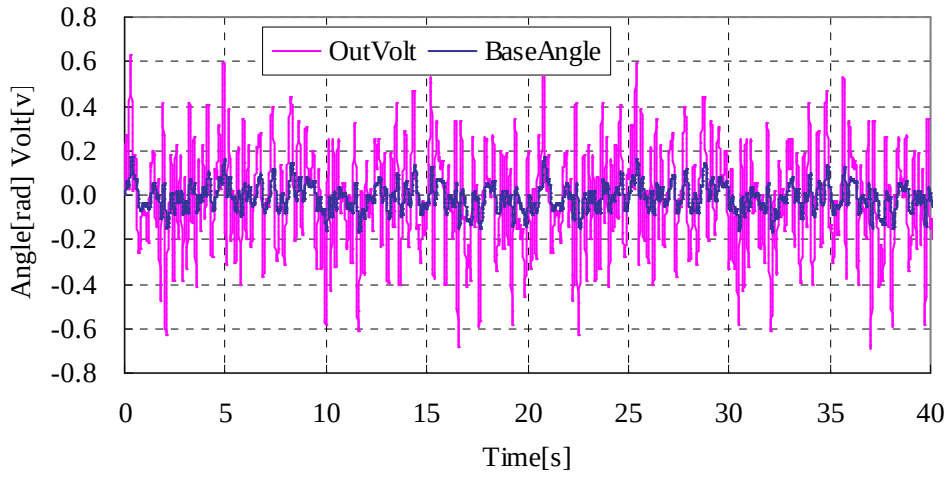


Fig.3-11 Input-Output Data (Band noise 0.0 to 5.0 [Hz])

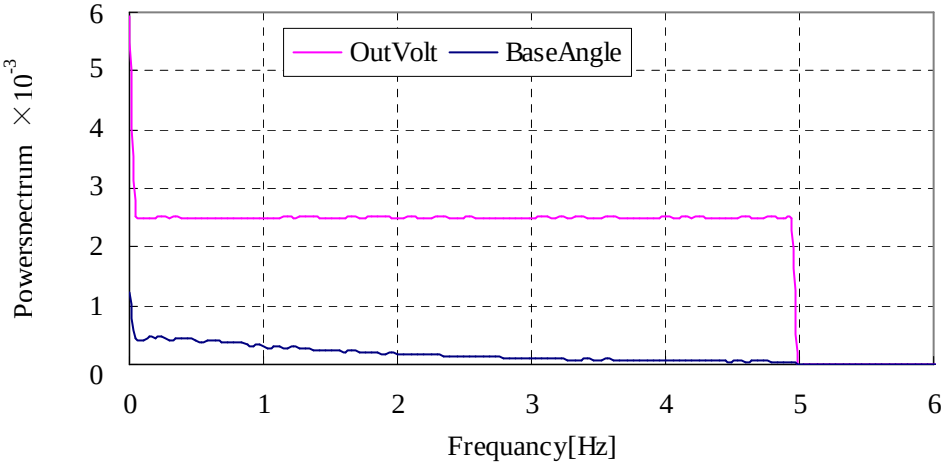


Fig.3-12 Input-Output FFT (Band noise 0.0 to 5.0 [Hz])

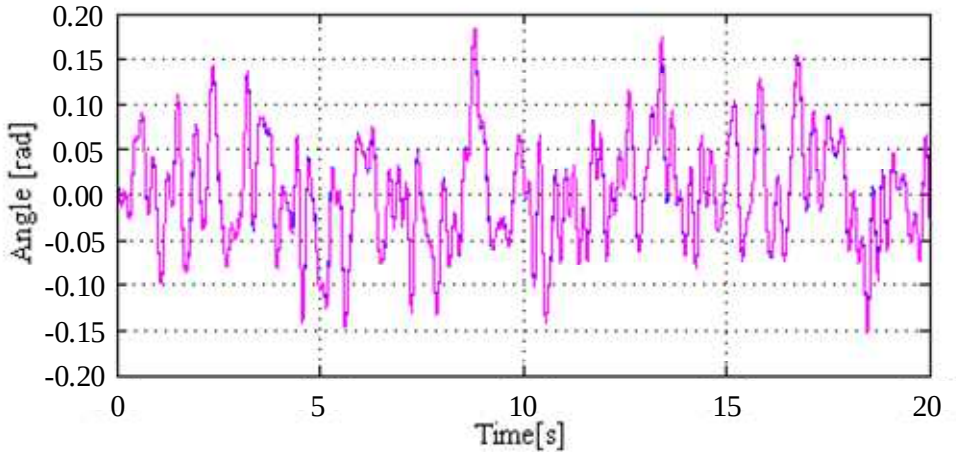


Fig.3-13 Plant output and model output (Band noise 0.0 to 5.0 [Hz])

次に Fig.3-14 に示すように, 作成したモデルに対して, 周波数の異なる 3 つの入力（1.5Hz, 3.0Hz, 20.0Hz）を与え，モデル出力と実際の出力を比較した結果を Fig.3-15～Fig.3-17 に示す．グラフからもわかるように，モデル化に使用した周波数帯の範囲内（1.5Hz, 3.0Hz）の入力データにおいては，モデル出力と実際の出力が非常によく合っているのに対して，モデル作成時の周波数帯を超える 20Hz のデータには 1.5Hz,3.0Hz のデータと比較すると，出力のずれが大きい（うまく適応できていない）ことがわかる．これらは人間の学習に例えると，習ったこと以外のことは答えることができない，ということを意味する．この習ったこと以外のデータに対する適応性を汎化性と言う．汎化性は機械学習における重要かつ本質的な問題と言える．

ここで汎化性について，更に深彫りする．上述の例では，モデル化に用いた周波数外の 20Hz の入力データにはうまく適応できなかった．そこでこれにも対応できるようにモデル化に用いる教師データに 20Hz の周波数も含む入力データを用いることを考える．

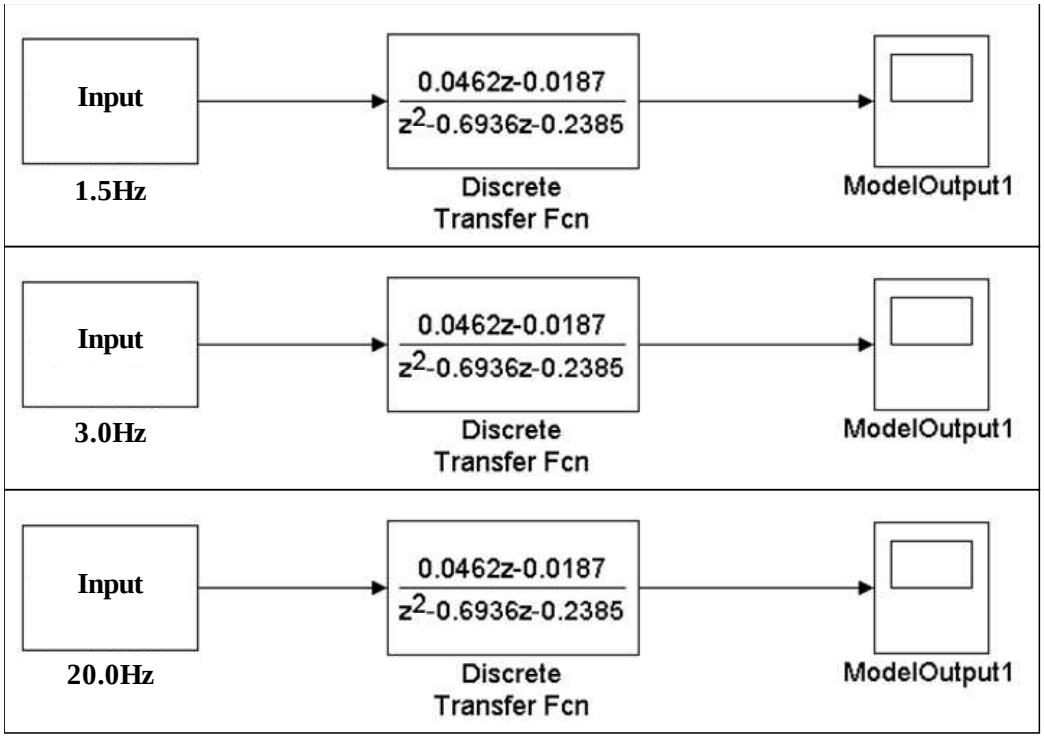


Fig.3-14 Simulation of model output

1. u(k)=0.3sin(3πt)　(1.5[Hz])

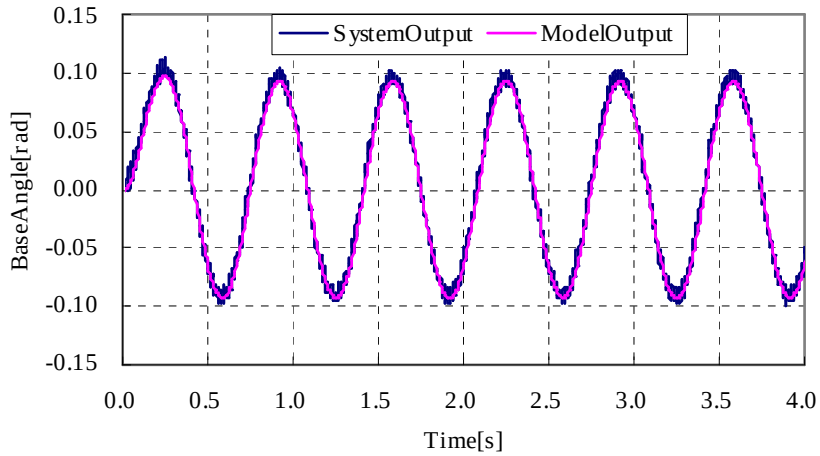


Fig.3-15 Plant output and model output (1.5[Hz])

2. u(k)=0.3sin(6πt)　(3.0[Hz])

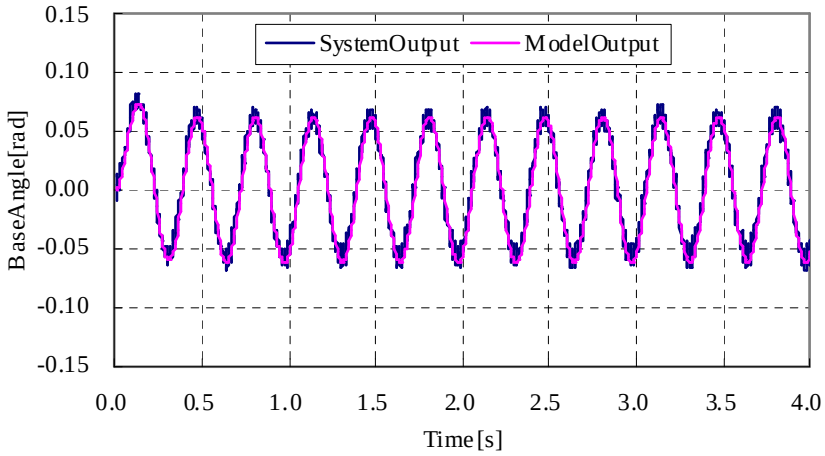


Fig.3-16 Plant output and model output (3.0[Hz])

3. u(k)=0.3sin(40πt)　(20.0[Hz])

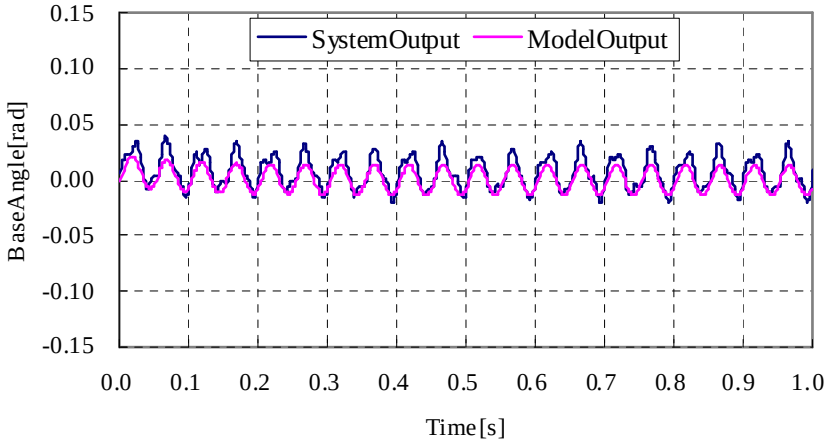


Fig.3-17 Plant output and model output (20.0[Hz])

Fig.3-18，Fig.3-19 に 0.0～25.0[Hz]の周波数を均一に含む入力，及び Fig.3-20 にモデリング結果を示す．Fig.3-20 を見ると，周波数帯を広くしたことにより，Fig.3-13 で示した 0.0～5.0[Hz]データでのモデリング結果よりもモデル化誤差が増えているのがわかる．もちろんこれが許容される場合もあるし，許容されない場合もあるが，一般には教師データに含まれる情報量が増えるほど，作成されるモデルは平均的なモデルとなり精度が下がる．人間の学習で例えると，広範囲のことを一度に覚えるのは難しいということである．ここに機械学習の本質的な難しさがある．

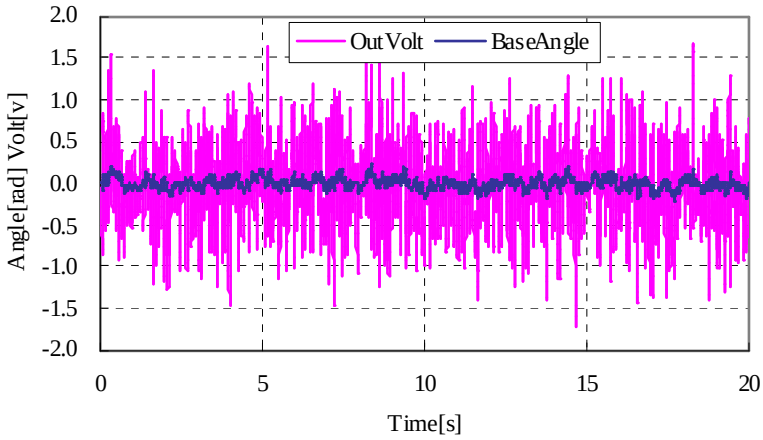


Fig.3-18 Input-Output Data (Band noise 0.0 to 25.0 [Hz])

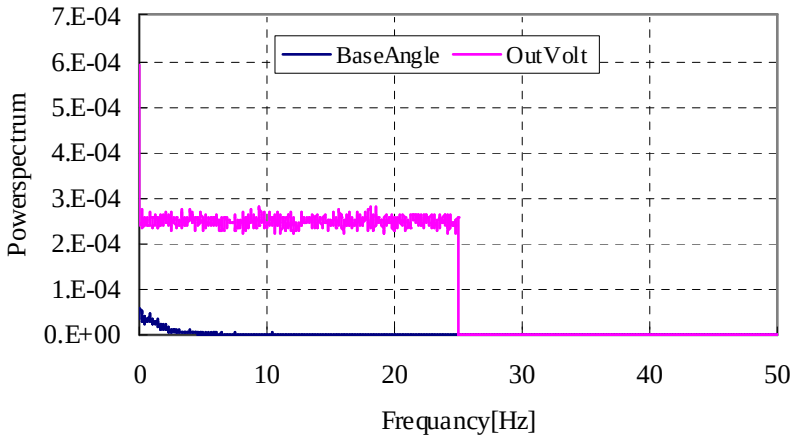


Fig.3-19 Input-Output FFT (Band noise 0.0 to 25.0 [Hz])

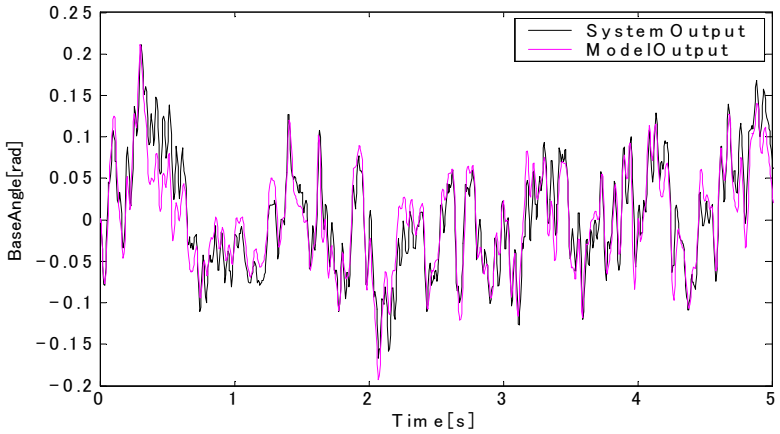


Fig.3-20 Input-Output FFT (Band noise 0.0 to 25.0 [Hz])

さて，ここで本タスクにおける機械学習の目的を再確認する．3・4・1 で述べたように通常の PID 制御では以下の 2 つの課題を有した．

①比例，微分，積分ゲインの調整が膨大なトライアンドエラーにより決定している
②全体の大きな振動の収束を速める為にハイゲインにすると，定常時の残留振動が発生する

これらの課題を解決するコントローラ（制御器）を学習するということがロボットに課せられた目的であった．この為にはコントローラを設計するのに必要なモデルを機械が自律的に獲得する必要があるのだが, 有用なモデルを獲得する上では, 人間が教示する余地があることを述べた．

しかしここで機械学習の一つのジレンマが生じる．機械に前提知識を与えれば与えるほど，元々の目的の「コントローラの自動最適化」という思想から離れていく．PID 制御のゲイン調整の負荷が，学習制御のモデリングの教示負荷に置き換わっては本末転倒である．そこで機械学習におけるこのジレンマを解消するために，次節ではモデリングとコントローラを機械が自ら改良していく繰り返し学習法を提案し, 機械と人間の望ましいインタラクションのあり方を議論する．

3･7　N.N.によるモデル/コントローラ繰り返し学習法

機械が所望のコントローラを学習する為にはモデルが必要だが，高いパフォーマンスを発揮するには高精度なモデルが要求される．しかし，汎化性の問題があり，広範囲に適応できるモデルを作成すると，精度の低い平均的なモデルになってしまうことを述べた．そこで，モデルを作成する教師データとして，実際に人間が設計してフィードバック制御を行った時の入出力データを用いることを考える．これは制御時の時系列データを教師信号として用いることで，実際の制御に近い帯域のデータで学習が出来るため，探索する必要の無い周波数帯のデータが不要となり効率的に学習が行われると考えたためである．しかし，一方で学習を行うロボットの探索範囲にはある程度の柔軟性を与えたい．そこで，制御としてはまだまだ改善の余地のある，適度に制御された P 制御時のデータを教師データとしてモデルを作成することにした．

Fig.3-21 に教師データを示す．今回与えた P 制御の比例ゲインは時間をかけて人間が最適に調整したものではなく，数回の試行で人間が調整したものであり，人間が“まずまず良好な制御である”と主観的に判断したデータである．(比例ゲイン=3.0)

この制御時の入出力データを教師信号としてロボットはモデリングを行い，コントローラを学習する為のシミュレータを作成する．具体的には Fig.3-22 のような NN を構成しコントローラの学習をすすめた．N.N.については第 2 章で述べた為，詳細を割愛するが，層数と入力，出力の次数を選定する必要がある．シミュレータ作成のためのシステム同定のモデル構造は入力 4（2 次遅れ系），出力 1 の 3 層の N.N.とした．　3・5・3 で述べたようにモデルの構造は重要な教示情報の一つであり，この N.N.の**ネットワーク構造**も**教示**の一つと言えるが，入力数に関しては，モータ＋フレキシブルアームの動特性は 2 次程度で十分であろう，という人間の常識的な選定と，ニューラルネットワークの構造としては非常に一般的な 3 層のネットワークを用いることで，人間の教示量の負荷を最低限に留めた．

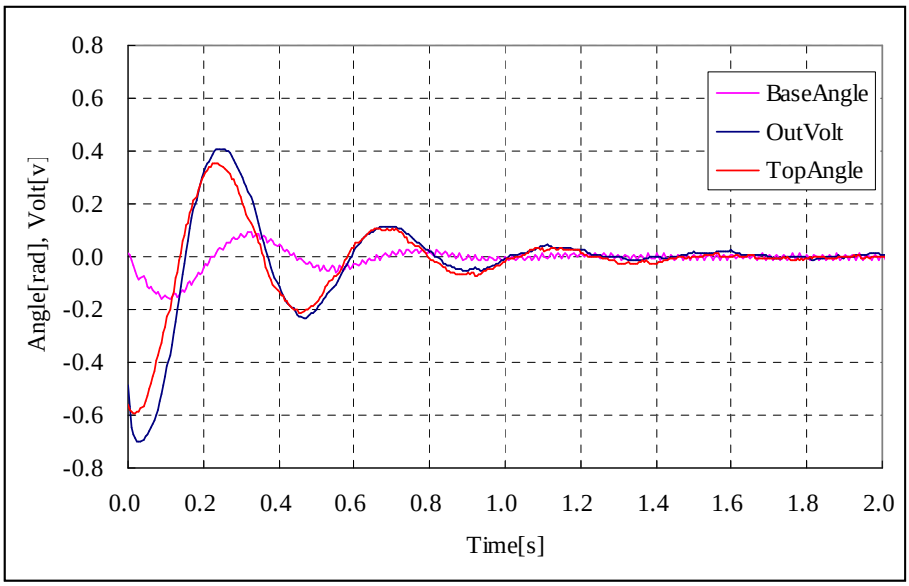


Fig.3-21 Input-Output Data (P Control)

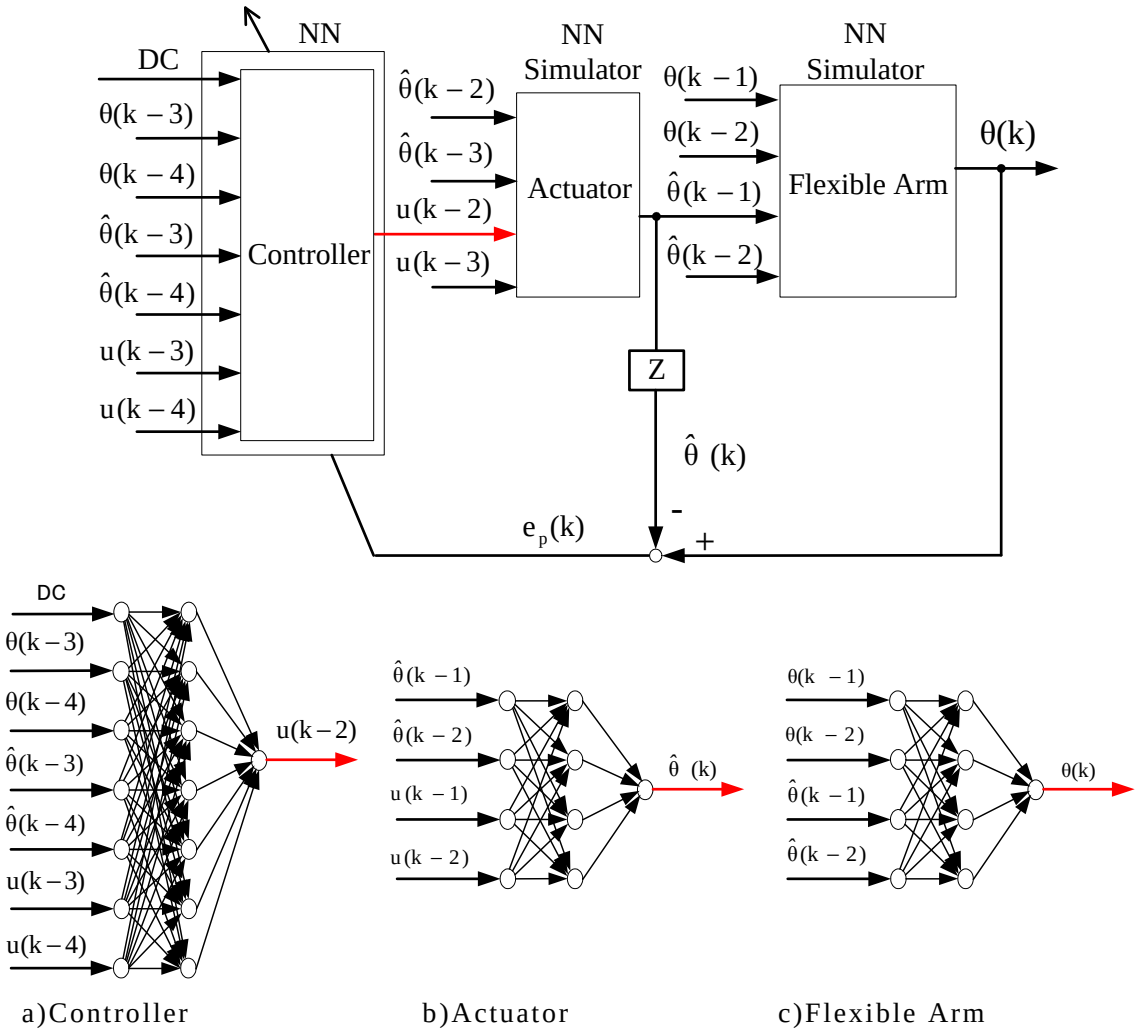


Fig.3-22 Neural Network Control

次にコントローラの学習方法について述べる．

コントローラが学習を進める際の評価関数は先端角度と根元角度の差分 $e_p(k) = \theta(k) - \hat{\theta}(k)$ であり，これを最小化するように学習を進める．

本手法では，ニューラルネットワークの汎化性の限界から，比較的良い制御性能が得られている P 制御時の実験結果を用いてシミュレータを作成し，それを基にコントローラを学習するが，シミュレータの動特性の表現能力の限界が，制御性能に影響を及ぼすものと考えられる．このため，本手法ではこの学習の不十分さを解決するために，学習で得られたより制御性能の良いコントローラを用いて制御した実システムのデータを用いシミュレータを作成し，再度そのシミュレータを用いてコントローラの学習を進めるという Off Line によるモデル/コントローラ繰り返し学習法を考案した．そのフローチャートを Fig.3-23 に示す．

ロボットは①の人間からの教示を起点として，②～⑤を繰り返し行うことで，モデル（シミュレータ）とコントローラを順次改善する自律性の高い学習手法となっている．②～⑤の 1 ループを繰り返し回数と呼ぶことにする．なおコントローラの構造もモデルと同様に一般的な 3 層の NN とした．

Fig.3-24 に学習結果を示すが，繰り返し回数が増加するとともに収束性能が改善されているのがわかる．

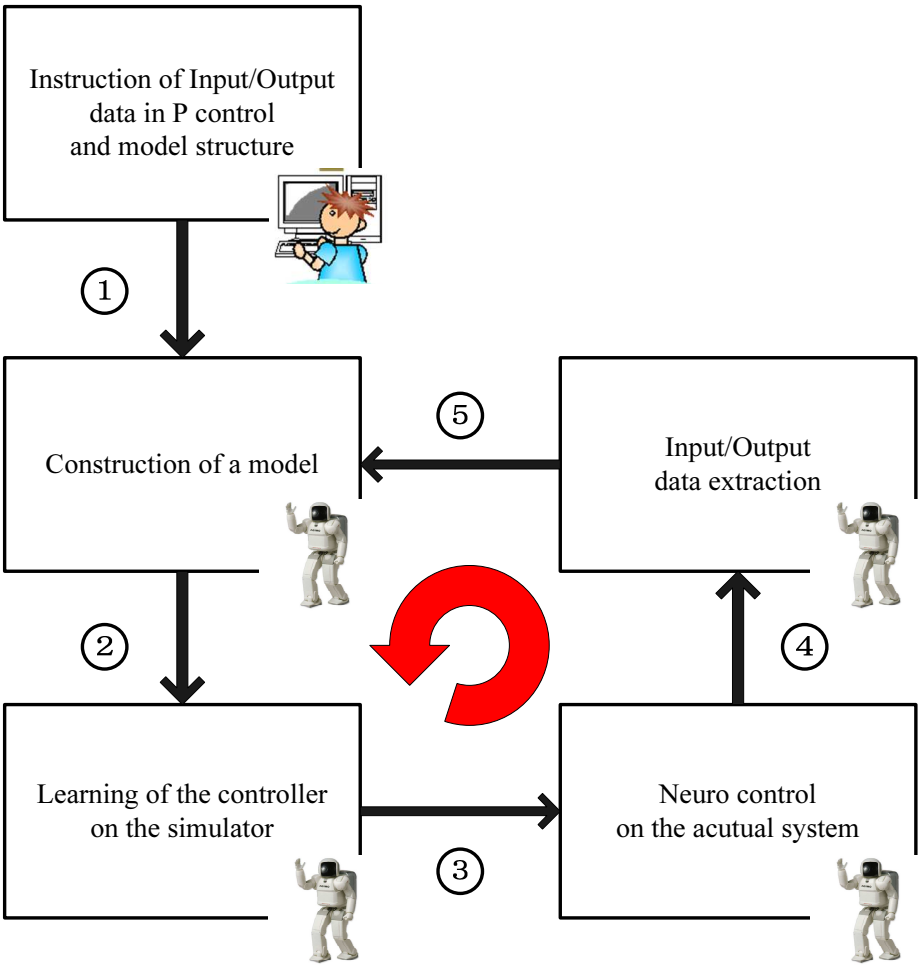


Fig.3-23 Flow Chart of Trial Learning Control

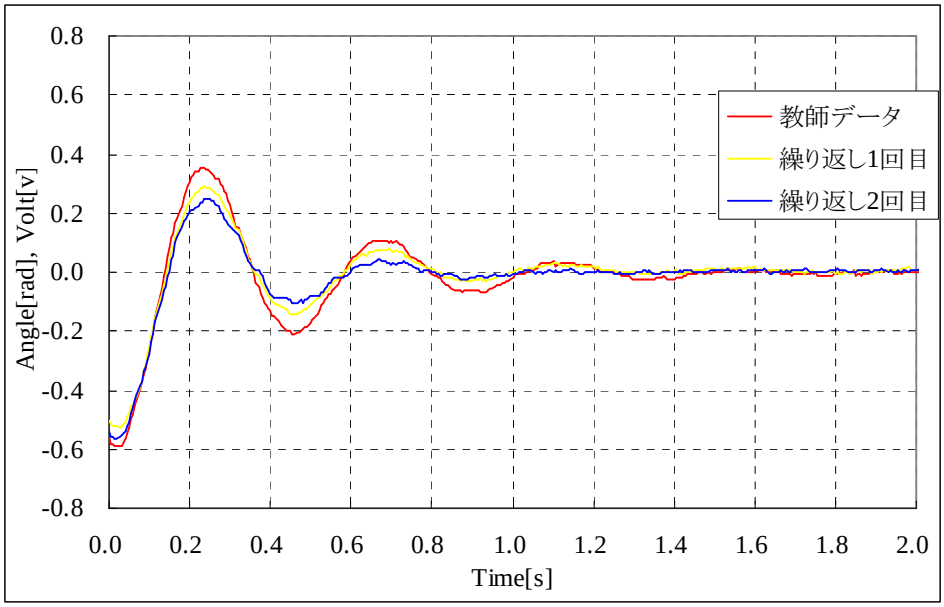


Fig.3-24 Learning result

3･8　人間と機械の協調学習（役割分担）

前節までに述べた繰り返し学習法では，機械は人間の最小限の教示を基に，よりパフォーマンスの高いコントローラの学習を行ったと言える．人間は制御対象モデルの大まかな構造判断や，コントローラゲインの簡易的な調整など，大枠を判断する作業は得意であるが，パラメータを最適化する作業は苦手である．（不可能では無いが，時間（コスト）がかかる）

一方で，機械は最適化問題を解くことは得意であるが，要所を抑えた大まかな探索が苦手であり，探索空間の絞込みには人間の手助けが必要となる．このように人間と機械には得意，不得意がある為，機械学習においては人間と機械の適切な協調が必要となる．

そこで機械学習における人間と機械の望ましい協調関係（役割分担）として Fig.3-25 に示す協調学習スキームを提案する．

0. 許容される制御目標値，学習時間を人間が設定する．
① 人間が主観的な判断に基づき，最低限の教示を機械に与える．
② 教示情報を基に機械がモデルを学習する．
③ モデルを基に機械がコントローラを学習する．
④ 学習したコントローラで機械が制御を行う．
⑤ 制御時の入出力データを用いて，機械がモデルを再度学習する
⑥ ②～⑤を繰り返し，学習が収束したところで，人間が結果を判断する．
⑦ 結果が OK であれば協調学習を終了，結果が NG であれば，追加の教示を行う．

機械への前提知識（教示）として，人間が機械へどのような情報を与えるか？その内容や優先度は対象システムによっても異なり，非常に複雑かつ難しい問題であることから，**「確かなところはわからない」**という前提に立つのが現実的であり実用的である．そこで人間の教示負荷が少な

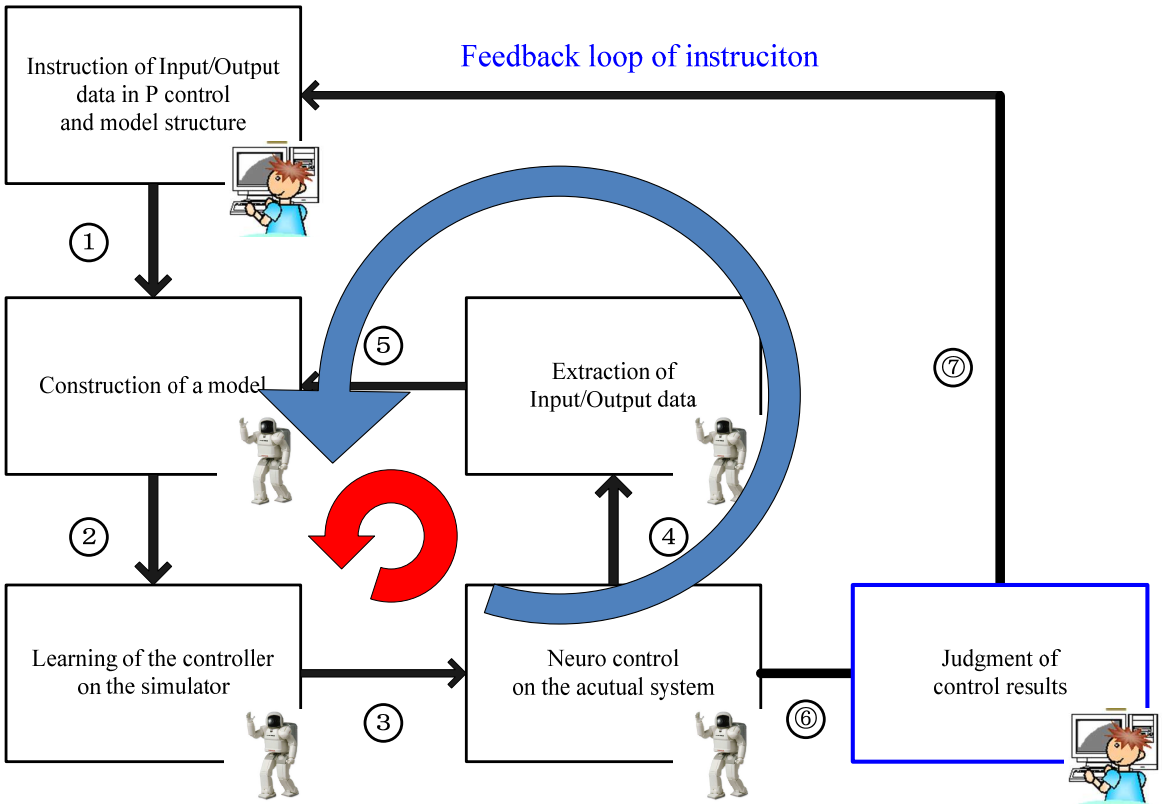


Fig.3-25 Cooperative learning scheme

い前提知識を与えることで教示の検討負荷を低減する．機械は人間からの教示情報を基に，学習により最適化を行い，人間は機械が行った学習結果を観察する．大きく学習結果が改善されれば，それは与えた教示が機械に対して有益な情報であったことを意味し，人間もまた機械の学習の変化具合を見て，機械の特性を外から理解することが出来る．学習結果に満足いかなければ，人間は新たな教示を行い，再度機械学習を行う．つまり，機械の最適化ループと，人間の教示ループの 2 つのループに基づき協調学習を行う．機械学習においては「人間の最低限の教示に基づく機械の自律的学習」が求められているという点では本協調学習手法は合理的な戦略と言える．

ここで 3・3 で述べた PID 制御のような非学習制御，通常の学習制御，及び本論文で提案した協調学習の違いを Fig.3-26 に示す．

非学習制御の一般的な設計の流れとしては，はじめに人間が制御対象のモデル構造を決定し，その後モデルパラメータの推定を行い，制御対象をモデル化する．次にモデルに対してコントローラ構造を設計し，最終的にコントローラのパラメータ調整を行う．一連の流れは場合によっては一部省略されることもあるが，これらは全て人間によって行われる．ここで注目すべきは，人間（設計者）と機械（制御対象）の接点は最終的にはコントローラパラメータ vs 制御結果にある点である．例えば PID 制御では，P,I,D のゲインを調整し制御結果を観察し，再度パラメータを調整し制御結果を判断する．もちろん制御結果がパラメータ調整だけでは満足いかない場合は，コントローラ構造の設計やモデル設計など上位に設計の観点をシフトさせるが，一般的にはコントローラパラメータ調整 vs 制御結果が人間と機械の最大の接点となる．

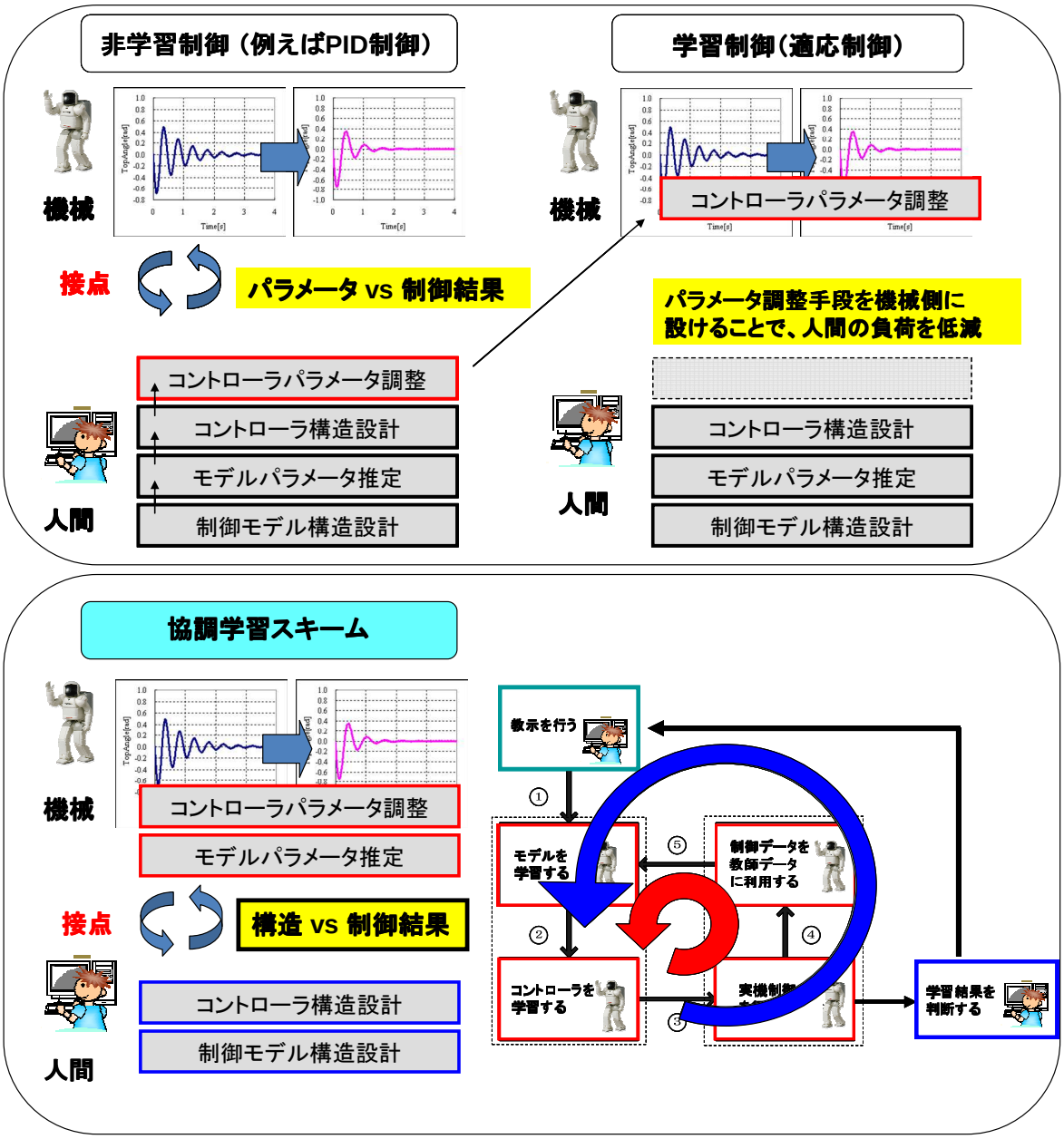


Fig.3-26 Cooperative learning scheme

学習制御ではコントローラパラメータ調整を機械側が行うことで，人間の設計負荷を低減させている．（学習手法によってはモデルパラメータの推定も機械が自動調整する場合もある）

一方，本論文で提案する協調学習では，人間はモデル，及びコントローラの**構造**を教示し，機械は繰り返し学習法に基づき**パラメータ**を調整する．つまり**人間の構造教示ループ**と，**機械のパラメータ調整ループ**の**役割を明確に分離**していることになる．

先にも述べたように，人間はタスクの大枠を把握し，教示することが得意であるから，協調学習においてはモデルやコントローラの構造教示に注力すべきであると考える．（Initial Guess）

一方機械は人間からの教示情報に基づき，与えられた評価関数のもと，パラメータを最適化することが得意であるから，パラメータ調整の役割を任せるべきであると考える．（Parameter Optimization）　このように機械学習における人間と機械の役割分担を，それぞれの長所・短所に基づき明確に規定することで，両者の望ましい関係を築くことが出来る．

なおここまでに，人間が適度に調整した P 制御時の教師データとモデル構造，コントローラ構造を大まかに教示することで, 機械がそれを更に改善していった例を示した. 第一章 序論にて**「機械は人間を越えることができるか？」**という哲学的な問いかけに対して No と記述した．これは，評価関数自体を機械は学習することが出来ないので，評価関数を決定する権限を与えられた上位の人間を概念的に超えることはできないという意味であった．しかし，今回機械学習により得られたコントローラは人間が教示した P 制御よりも良好な制御結果となっており，人間を超えている．つまり**「機械は人間を越えることができるか？」**という問いかけは定義が曖昧であり，人間が教示した教師データ以上の性能を機械が獲得することができるか？という問いかけに対しては Yes と言うことになる．

3･9　機械の学習過程の推察

前節までは機械学習過程での人間と機械の望ましい関わり方について述べたが，この学習過程で機械がどのようなコントローラを学習したかは興味深い為，本論とは外れるが深彫りする．

ニューラルネットワークはその構造上，内部がブラックボックスであり制御においてはどのような性質をもつコントローラが学習されたかが外部からは分かりづらい．そこで Fig.3-27 に示すように，先に有効性が示された繰り返し学習が行われたニューロコントローラに P 制御を併用し制御を行うことで制御性能がどのように変化するかを見て，機械の学習過程を推察・理解する．

Fig.3-28 に結果を示すが， P 制御を加えることで収束性能が大幅に向上しておりオーバーシュートも最小限に抑制されていることが分かる．このことから推察されることは本学習でニューラルネットワークが獲得したのは，オーバーシュートの抑制による D 制御的効果が主体であったと考えられる．本手法では，適度に調整された P 制御の制御結果を用いてシミュレータを作成し，それを基にコントローラを作成したのであるが，収束性の良い制御データを用いるとデータ数が少なくなり，システム同定精度が悪化する．このシステムの同定精度を向上させるために，若干，

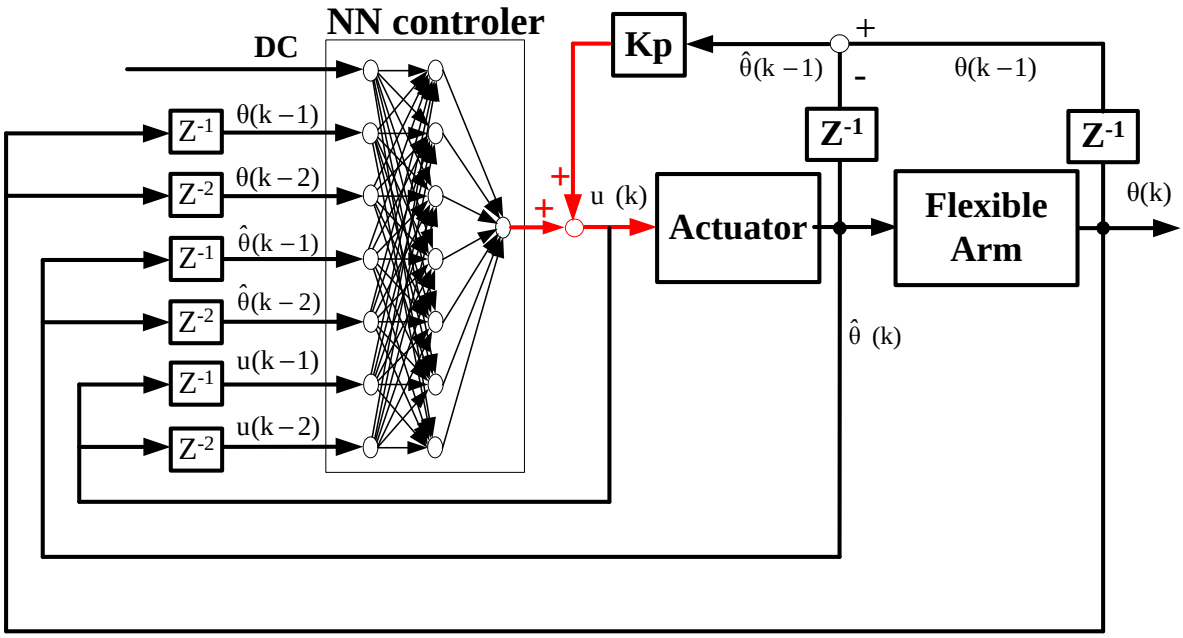
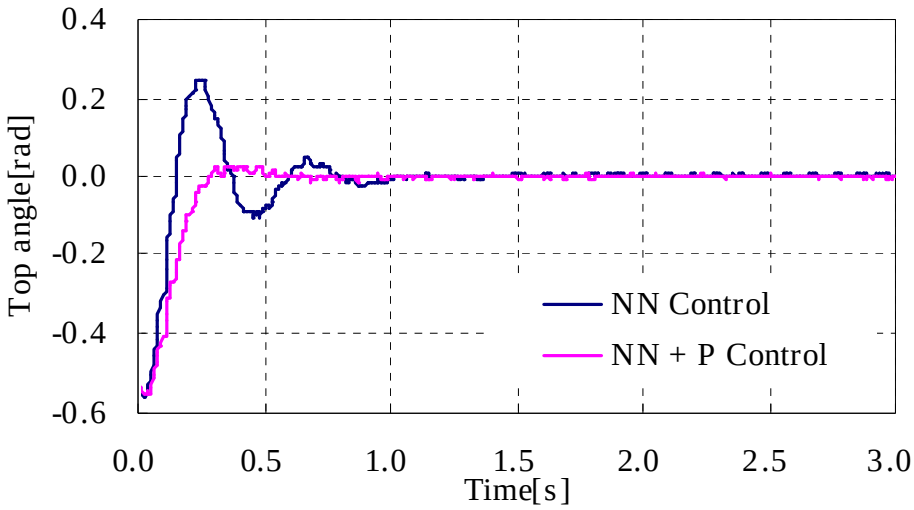
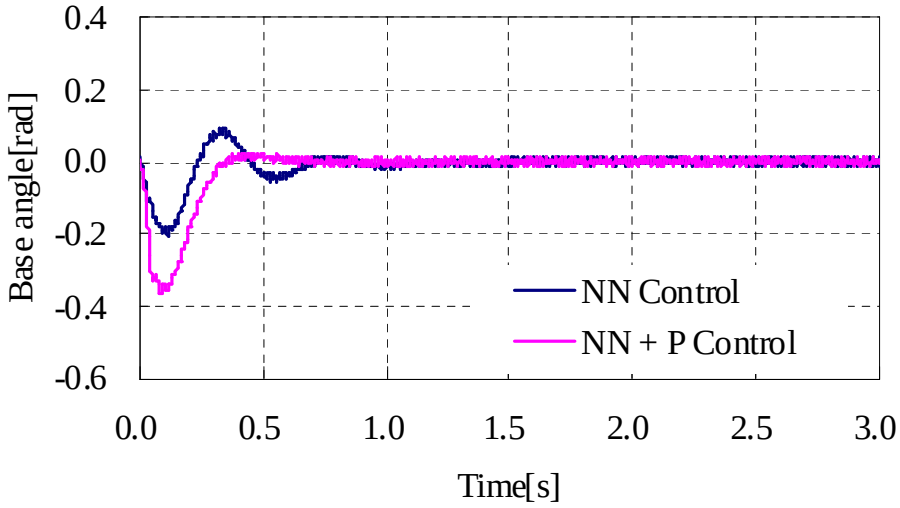


Fig.3-27 NN Control with P Control





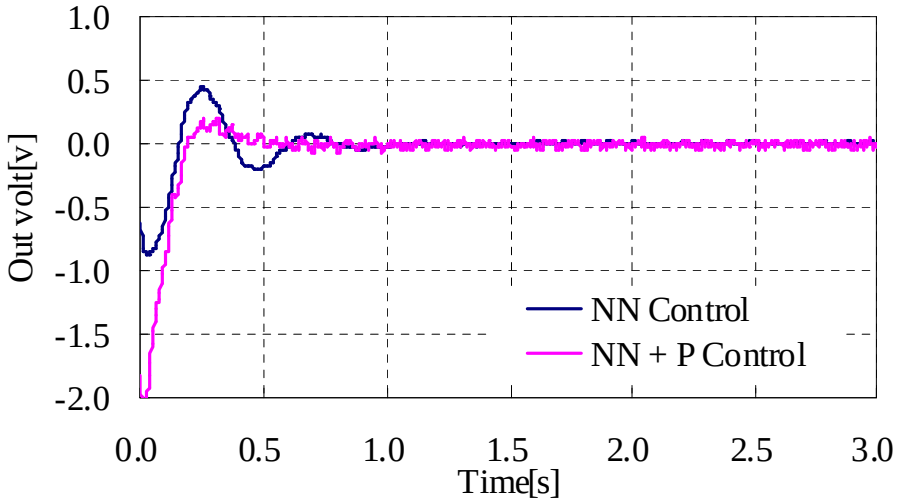


Fig.3-28 Learning Result

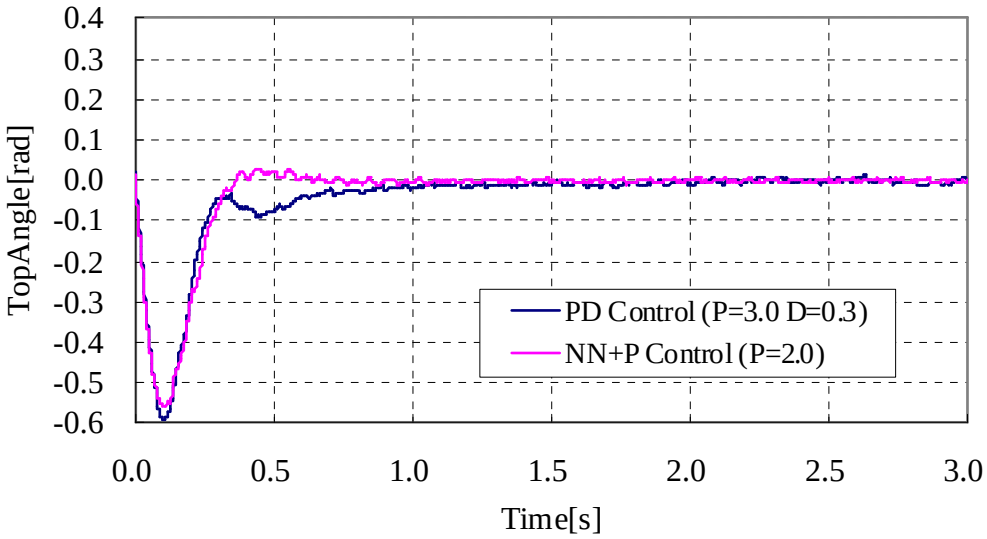


Fig.3-29 Learning Result

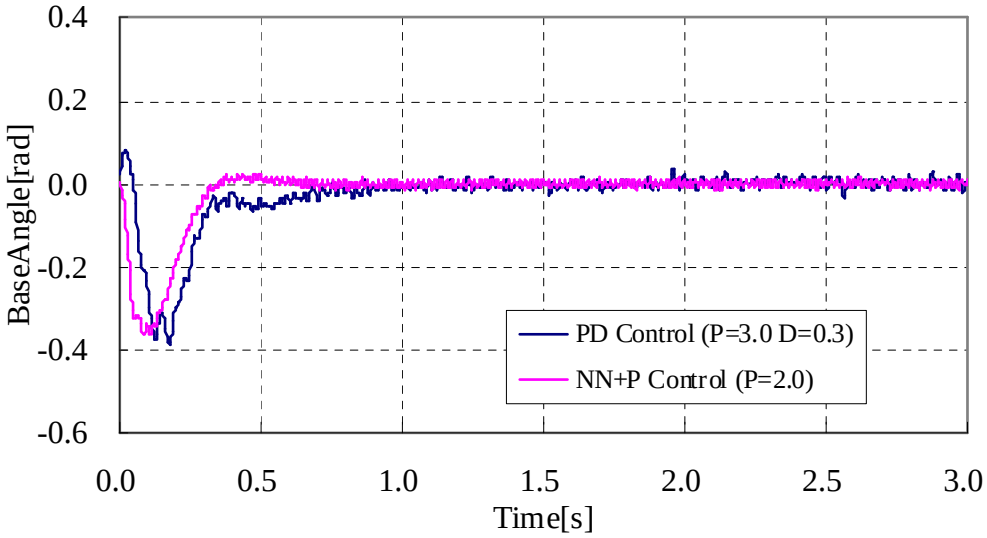


Fig.3-30 Learning Result

収束性の悪いデータを用いてシミュレータを作成している．この結果，P 制御則についてはベストの制御則が得られず，併用手法において最高の性能が得られたと推定できる．

次に今回機械学習で得られたコントローラと，人間が調整した最適な PID 制御結果を比較する．

Fig.3-29 及び Fig.3-30 に比較結果を示すが，先端角度，根元角度ともに P 制御併用型非線形 NN 制御は PD 制御と同等以上の制御性能を有していることがわかる．この点から 3・4・1 で示した PID 制御の課題①はクリアしている．

次に課題②である残留振動が増幅されていないかを確認した．初めに出力している指令電圧を比較する．Fig.3-31 に示すように PD 制御がノイズ成分を受けて，収束した後も高周波の指令電圧を出力し続けているのに対し，P 制御併用型非線形 NN 制御ではフィルタリング効果が働き高周波の信号はカットしているのがわかる．これを受けて，先端角，根元角の残留振動を比較する．

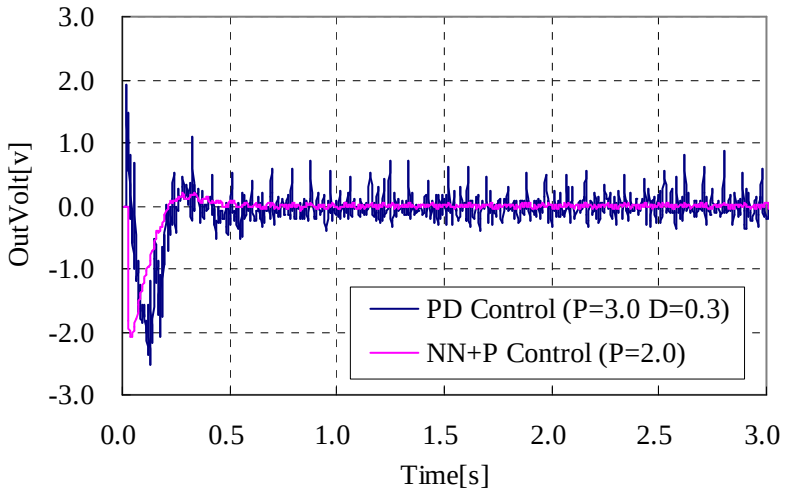


Fig.3-31 Learning Result

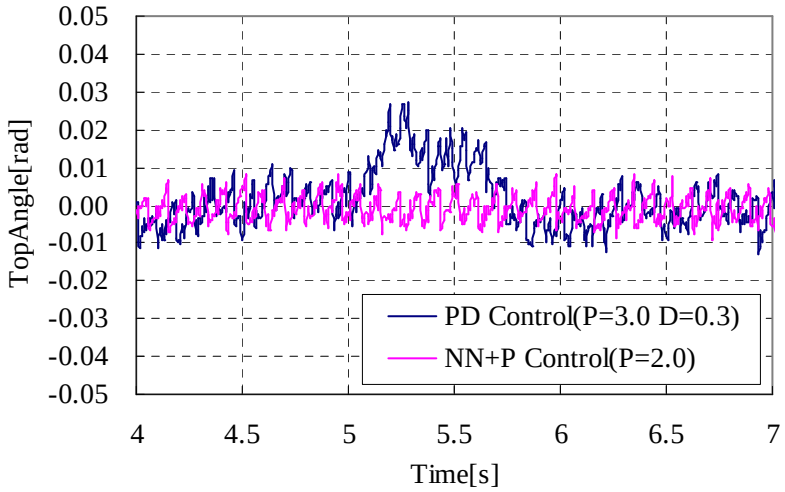


Fig.3-32 Learning Result

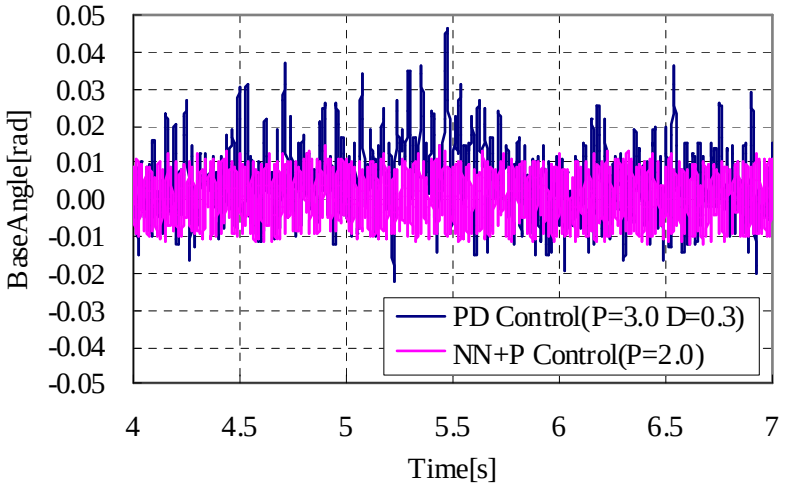


Fig.3-33 Learning Result

Fig.3-32 及び Fig.3-33 を見ると，NN がフィルタリング効果の役割を果たし，先端角，根元角の残留振動を増幅しないという結果になった．このことから構成した NN は速度成分を補償するだけでなく，フィルタの役割も同時に果たすコントローラとなっていることが明らかになった．よって前述した①②の問題を解消するコントローラを獲得したということになる．

3･10 教示の主観性

本タスクはフレキシブルアームロボットが振動を抑制するコントローラを学習により獲得することであった．本論文では機械に対する前提情報を極力排除し，機械自らがモデルとコントローラを逐次改良する，繰り返し学習法を提案した．このコントローラを機械が学習するために人間が教示したことは，「人間が適度に調整した P 制御の教師データ」と，「常識的なモデル，コントローラ構造」のみである．人間が行う比例ゲインの調整はわずかの時間でよく，それを基に機械がより最適なコントローラを獲得したことになる．つまり**人間が得意とする大枠を定めるスキル**と，**機械が得意とする細かな最適化のスキル**がミックスされた形となる．

ここで次章以降の議論のきっかけとして，教示の主観性について述べる．今回人間が与えた教示データは「人間が適度に調整した P 制御の教師データ」と「常識的なモデル，コントローラ構造」であり，この判断には人間の**主観性**が多分に含まれている．

つまり「人間が適度に調整した・・」の**“適度に”**は，何をもって適度なのか？「常識的なモデル・・」の**“常識的な”**は，何をもって常識的なのか？という判断は実に主観的なものであり，人間の経験や知識レベルなどにより異なる．この人間の主観的判断が機械の学習にどのような影響を与えるのかは大変興味深い問題である．

そこで機械学習における主観性の問題について第 6 章，第 7 章で議論する．

3･11 まとめ

本章ではフレキシブルアームの制振問題をタスクとして，ニューラルネットワークを用いた機械学習を適用した．その中で人間の持つ優れた能力（大枠を判断する能力）と，機械の持つ優れた能力（最適化問題を処理する能力）について述べた．またモデル/コントローラ繰り返し学習法を提案し，人間と機械が協調するフローを述べた．

一般的にニューラルネットワークは教師あり学習法として，非線形性の強いシステムのモデルジェネレータとしての役割を期待されることが多いが，本手法ではニューラルネットワークを一つの学習ツールとして，人間と機械の違い，協調のあり方についての構想を述べたものであり，他の機械学習アルゴリズムでもエッセンスを用いることが可能である．

「教師あり学習」,「教師なし学習」と言う機械学習の区分方法はアルゴリズム的分類としての利便性はあるが，本質的には機械学習において人間と機械がどのように関わるか，どのような得意・不得意をカバーしあうのかということを議論することが重要であり，本章ではこの観点から両者の望ましい関係を明らかにした．

第4章
強化学習概説と実験システム

4･1 概説

本章では強化学習の理論及び実験システムについて述べる．理論の説明では特に強化学習の特徴である，将来の報酬を考慮に入れた『価値』というものに主に焦点を当てて説明する．次に，これらが実際にどのように定式化されるかについて述べた後，強化学習で最も広く使われている TD 学習について説明する．その中でも TD 学習の代表的な学習アルゴリズムのうち特に応用例が多く，その有効性が確かめられている Q-Learning および Actor-Critic について説明する．

4･2 強化学習の概要

強化学習とは生物の適応過程を工学的観点からモデル化した枠組みである．例えば人間の赤ん坊が学習をする場合を考えてみる．赤ん坊は生まれて間もなく，泣き声をあげ，その後よちよち歩きをはじめ，しばらく経つと自ら立ち上がり二足で歩くようになる．この成長の過程ではどのようなことが行われているだろうか？赤ん坊が行動する空間には，母親，父親であったり，部屋に置いてある玩具であったり，赤ん坊に影響を与える何らかの人や物がある．赤ん坊はこれらからの何らかの影響を受けながら，自ら行動を獲得していくのである．例えば，赤ん坊がよちよち歩きをしていて，ある時たまたま立ち上がったとする．母親はそれに対して，「よしよし」と誉めてやることにより，赤ん坊はそれが良い行動であったことを認識し，成長していく．

このような環境との相互作用により，行動主体となるものが自ら行動を獲得していくプロセスを模倣し，アルゴリズム化した枠組みを強化学習という．

強化学習では，赤ん坊のような行動主体となるものを『エージェント』，母親のような褒美を与えてやる存在を『環境』，環境から与える褒美を『報酬』としてモデル化する．Fig.4-1 にその枠組みを示すが，エージェントは環境（制御対象）の状態を観測し，その状態に対応した行動を行う．エージェントの行動によって環境の状態が遷移し，エージェントはその遷移に応じた報酬を得る．つまり，エージェントは状態観測→行動→状態遷移→報酬の取得を繰り返し，学習が行われていく．このように強化学習は元々人間などの成長過程を模倣しているのだが，1990 年代に機械学習や人工知能の分野で，またオペレーションズ・リサーチ，心理学や神経科学などの分野でも人気となり，現在学習研究の主流の一つとなっている．強化学習の基礎的な理論については過去に多くの検討がなされている[(43)~(83)].

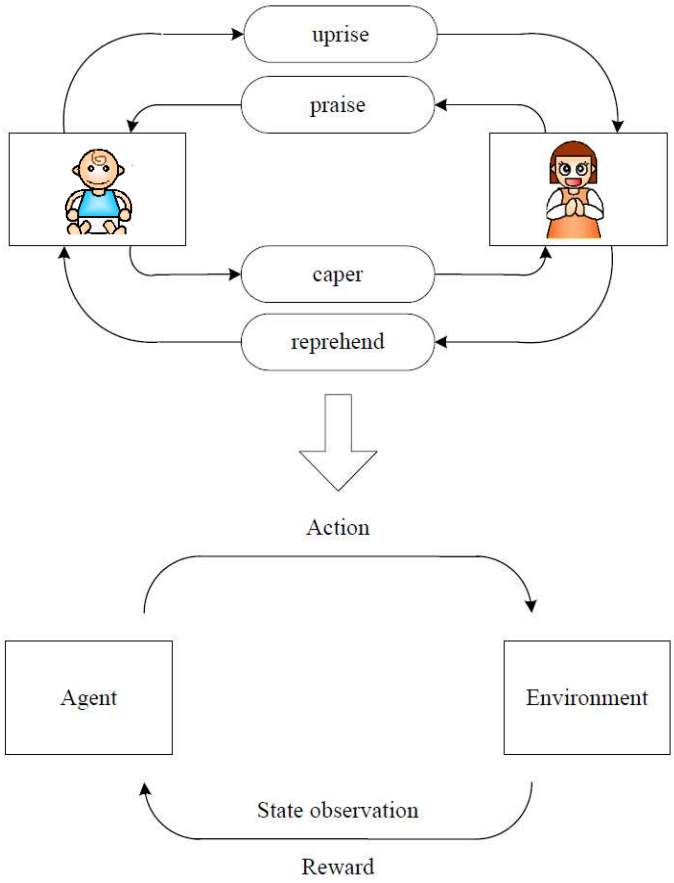


Fig.4-1　Frame work of reinforcement learning

4･3　強化学習の特徴

本節では他の学習手法と違う，強化学習の特徴をいくつか述べる．

(1) 教師なし学習である

前述のように，赤ん坊は母親などから「この足をこう曲げなさい」と事細かに正解を教わるわけではなく，行動した結果に対して良いか悪いかの評価(ヒント)が与えられることから，正解を直接的に明示する教師のいない，教師なし学習の一つとして位置づけられ，ニューラルネットワークなどの教師あり学習とは異なる分類がなされる．

(2) 遅延報酬の処理が必要である

エージェントが何回かの状態遷移を繰り返した後に，初めて報酬を得ることができるような環境が数多く存在する．例えば，迷路を探索しゴールを探すタスクの場合，エージェントはゴールにたどりついた場合にのみ報酬を得ることができる．このような場合，行動を選択した直後に報酬を得ることができず，報酬が遅れてしまう．よって，行動を実行した直後の報酬をみるだけでは，エージェントはその行動が正しかったかどうかを判断できないという困難を伴う．これについては後に述べるが『価値』という指標を導入し評価する．

（3）探索と知識利用のトレードオフ問題

多くの報酬を得るために，強化学習エージェントは過去に試みた行動の中で，報酬を得るために効果的なものを優先的に選ばなくてはならない．ところが，このような動作を発見するためには，過去に試みたことのない行動も選択してみなくてはならない．つまり，エージェントは報酬を得るためにすでにもっている知識を利用し，将来的に行動選択を改善するためには探査も行わなくてはならない．ここで生じるジレンマは，探査も知識利用も与えられた作業の失敗なしに独自に遂行されることはないということである．エージェントはいろいろな行動を試し，その中で最良と考えられるものを徐々に見出していく必要がある．

以上のように強化学習には他の学習手法にはないいくつかの特徴を持っており，それらの特徴を活かすことで様々な利点が考えられる．以下に応用上期待されることを記す．

（1）適用範囲が非常に広い

適用先が問題に対する意思決定といった上位層レベルの問題[110]~[129]から，モータの出力角といった非常に下位層レベルの問題[130]~[160]まで幅広いことが上げられる．強化学習が広く使われるようになったのはその為であろう．例えば上位層レベルの問題においては浅田らがサッカーロボットにおける行動獲得に適用し，“静止する”，“ゴールに向かう”，“他のエージェントに向かう”といった大きな意思決定の学習に強化学習を用いている[129]．また Andrea らはロボットがケーキを焼く為に必要な“具材を運ぶ”，“卵を割る”といった行動選択の学習に強化学習を用いている[203], [204]．工学分野以外で言えば石井らのオセロの戦略に適用した例が有名であり，人間同等以上のパフォーマンスを発揮できることが知られている[110]．一方，下位層レベルの問題への適用としては銅谷らの研究が有名であり，生物型ロボットの動作獲得に適用している[130]．銅谷らの研究では生物の学習過程と強化学習アルゴリズムの関連について議論がなされ，様々な形態のロボットに適用された．また木村や伊藤らも実ロボットの前進行動獲得に適用し，強化学習が下位層レベルの学習にも有効であることが広く認識された[131], [132]．

（2）制御プログラミングの自動化・省力化

これまで示したように，強化学習では報酬という単一のスカラー量のみを与えることにより，エージェントの学習が行われる．一般に環境に不確実性や計測不能な未知のパラメータが存在すると，タスクの達成方法やゴールへの到達方法は設計者にとって自明ではない．よってロボットへタスクを遂行するための制御規則をプログラムすることは設計者にとって非常に手間や時間がかかる場合が多い．ところが，達成すべき目標を報酬によって指示することは，前記に比べれば遥かに簡単である．そのため，タスク遂行のためのプログラミングを強化学習で自動化することにより，設計者の負担軽減が期待できる．十分に優れた性能を持つ強化学習エージェントをコントローラとして１つだけ開発しておけば，あとはロボットの目的に応じて報酬の与え方だけを設計者が設定するだけで，あらゆる種類のロボット制御方法を同一のコントローラによって自動的に獲得できる．

（3）人間が想定する以上の解の発見

エージェントは試行錯誤を通じて学習するため，人間のエキスパートが得た解よりも優れた解を発見する可能性がある．特に不確実性（摩擦やガタ，振動，誤差など）や計測が困難な未知パラメータが多い場合，人間の常識では対処し切れないことが予想され，強化学習の効果が期待できる．エキスパートの制御規則を学習初期状態に設定してそれを改善する場合と，全くのゼロから学習を開始し，設計者にとっては意外な新しい解を発見する場合とが考えられる．

4･4　強化学習の研究動向

強化学習の従来研究の多くをまとめれば，三つの大きな潮流がある．

第一に強化学習アルゴリズムの研究である[39)~(83)]．強化学習の核となる考え方は「報酬」というスカラー量を元に「行動価値(いわば知識)」を学習することにあるが，そのアルゴリズムには様々な手法があり，過去に広く研究がなされてきた．特に広く使われているアルゴリズムとしては，TD 誤差指標[60)~(64)]を元に学習を進めるアルゴリズムとして，Q-Learning[68)~(72)]や Actor-Critic[73)~(76)]などが提唱されている．強化学習アルゴリズムの分野はすでに体系化されているものの，現在でも学習効率の良いアルゴリズムについて研究が進められている．

第二に，連続系，動的問題への拡張があげられる[84)~(93)]．通常強化学習はあるタスクに対して，行動パターンをいくつかに分割し，それらを組み合わせることでタスクを達成していく．この分割数が小さければ行動は単調なものとなり，タスクを達成させる面においても，支障が出ることになり，何より見た目に非常に面白みのかける行動となる．一方，行動パターンを大きく分割すれば細かい作業が可能である一方で，空間の情報量が膨大となり（組み合わせの爆発問題，あるいは次元の呪い問題と言う）それらを保持するメモリ量，学習時間といった面で現実的ではない．そこで，経験したことのない行動に対し，それらを補完・予測する汎化の問題が重要となり，連続系・動的問題への拡張の研究が自然と拡大していった．

第三としては強化学習の各種実問題への適用（アプリケーション）である[110)~(186)]．先にも述べたように，強化学習は上位層の問題から下位層の問題まで，実問題への適用例が非常に多い．最近ではコンピュータ処理能力の向上も手伝い，状態数の多いより複雑なタスクへの展開例も増えてきており，近年の研究の中心となっている．学習アルゴリズム自体はクラシカルなものでも，過去に展開が困難であった問題への適用が可能になり，その結果から新たな知見を得て研究が加速するなど，強化学習の研究は近年，再度注目を集めている．

4･5　強化学習の定式化

ここまで，強化学習がどのような成り立ちで模倣され，応用されるのかについて概念的に示し，従来研究について概要を述べた．本節では実際に問題に適用する上で，どのように定式化され利用されるのかについて述べる．

4･5･1　基本的な取り決め

先に述べたように，強化学習は環境との相互作用から学習して目標を達成する問題の枠組みである．学習と意思決定を行うエージェントと，環境は離散的な時間ステップ t=0,1,2,3,･････の各々について相互作用を行う．各時間ステップ t において，エージェントは何らかの環境の状態の表現 $s_t \in S$(S は可能な状態の集合)を受け取り，これに基づいて行動 $a_t \in A(s_t)$を選択する．1 時間ステップ後に，エージェントはその行動の結果として数値化された報酬 $r_{t+1} \in R$ を受け取り，新しい状態 s_{t+1}にいることを知る．

以上をまとめると，強化学習においては

現在の状態	s_t
行動	a_t
遷移先の状態	s_{t+1}
環境から受け取る報酬	r_t

という表現を多く用いる．

4･5･2　方策

先に示したように，エージェントは状態を観測して，それを元に行動を選択する．この状態から可能な行動を選択する確率の写像をエージェントの『方策』と呼び，π_t で表す．ここで，$\pi_t(s,a)$ は，もし $s_t = s$ ならば，$a_t = a$ となる確率を示す．例えば，左から右にしか動かないエージェントがいた場合，そのエージェントの方策はあらゆる状態 s に対して

$$\pi_t(s,\text{right}) = 1, \quad \pi_t(s,\text{left}) = 0$$

であり，左と右に同じ確率で動くようなエージェントの場合，そのエージェントの方策は

$$\pi_t(s,\text{right}) = 0.5, \quad \pi_t(s,\text{left}) = 0.5$$

となる．後に述べるが，強化学習は最終的に受け取る報酬の総和を最大にするように，方策を変更していく学習手法である．

4･5･3　収益

強化学習では，ある『状態』において何らかの『行動』をし，状態が変化すると，それに応じて環境から『報酬』が得られることを述べた．

一見，ある状態から次の状態に遷移した際に得られる報酬が大きい場合，その行動はよかったと考えられるが，例えば遷移した先の状態ではこれ以上どんなに動き回っても報酬がもらえないような場合は，先の行動が良かったとは一概には言えない．逆にある行動を取ってあまり報酬が得られなかったが，遷移した先の状態で，何らかの行動をした結果，非常に大きな報酬がもらえ

るような場合, 先の行動はよかったことになる. これらを図で示したものが, Fig.4-2 である. Fig.4-2 では，丸の中の数字を状態番号，直線を行動経路，線横にある数字を報酬として描いている．一見スタートの状態では左に進んで 1 の状態に移ったほうが報酬が多く貰えるが，遷移した先の 1 の状態ではその先どのように進んでも大きな負の値しかもらえない．一方スタートの状態で右に進んでいれば，はじめは - 5 という負の値を受けたが遷移した先の 2 の状態では，その先に大きな正の報酬を得られるとすれば，スタート地点では右に行った方が『良い行動』なのである．

このように強化学習では，目先の利益のみを評価するのではなく，ある一連の行動でトータルに得る報酬を最大化するように学習を行うことを考える．

以上を定式化することを考える．ある時刻 t の後のステップに受け取った報酬の系列を $r_{t+1}, r_{t+2}, r_{t+3}, \cdots$ とすると，トータルの報酬 R_t は

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_T \tag{4-1}$$

となる. このトータルの報酬 R_t を『収益』といい, 強化学習ではこれを最大化することを考える.

ここに式(4-1)で示した T は最終時間ステップを意味している．例えばオセロゲームでは，プレイヤーが何回かの選択の後，全ての盤面が埋め尽くされた時点でゲームが終了するし，迷路問題で言えば迷路のゴールに到達した時点でタスクが終了する．このように終了状態が存在するような場合，スタートからゴールまでの一連の行動を一つの固まりとしてとらえ,『エピソード』と呼び，そのようなタスクを『エピソードタスク』と呼ぶ．このような場合，各エピソードは終端状態という特殊な状態で終わり，その時刻の報酬を r_T と表わしている．

しかし全てのタスクがエピソードタスクというわけではない．例えば，本研究のようなロボットの前進行動の獲得などの場合，明確なゴールが存在するわけではなく，行動は無限回続く．このような終端状態が存在しないタスクを，先に示したエピソードタスクと対比して『連続タスク』と呼びこのようなタスクの場合 $T=\infty$ となるから式(4-1)の収益は

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots = \sum_{k=0}^{\infty} r_{t+k+1} \tag{4-2}$$

となり，収益が発散することから，これらを最大化することは意味を持たなくなる．そこでこのような連続タスクでは『割引』という概念を導入し，次式で示す減衰収益を最大化することを考えることにする.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{4-3}$$

ここに示した γ は割引率と呼ばれるパラメータで $(0 \le \gamma \le 1)$ である．つまり，エージェントが現時刻 t において考えたとき，将来的に得る報酬 $r_{t+2}, r_{t+3} \cdots$ と言うのは非常に不確かで，どれだけの報酬が貰えるのかは現時点では定かではないので，未来の報酬であればあるほど割り引いて見積もるという考え方である.

こうすることで報酬の系列の加算を無限回繰り返しても，その値は有限となり収益を評価することができる．式(4-3)においてもし割引率 γ を 0 とすれば収益

$$R_t = r_{t+1} + 0 + 0 + \cdots\cdots + 0 = r_{t+1} \tag{4-4}$$

となり，エージェントの目的は収益 R_t を最大にすることであるから，エージェントは目先の報酬 r_{t+1} (即時報酬)のみに関心を持つことになり，近視眼的な行動選択が行われるようになる．このよ

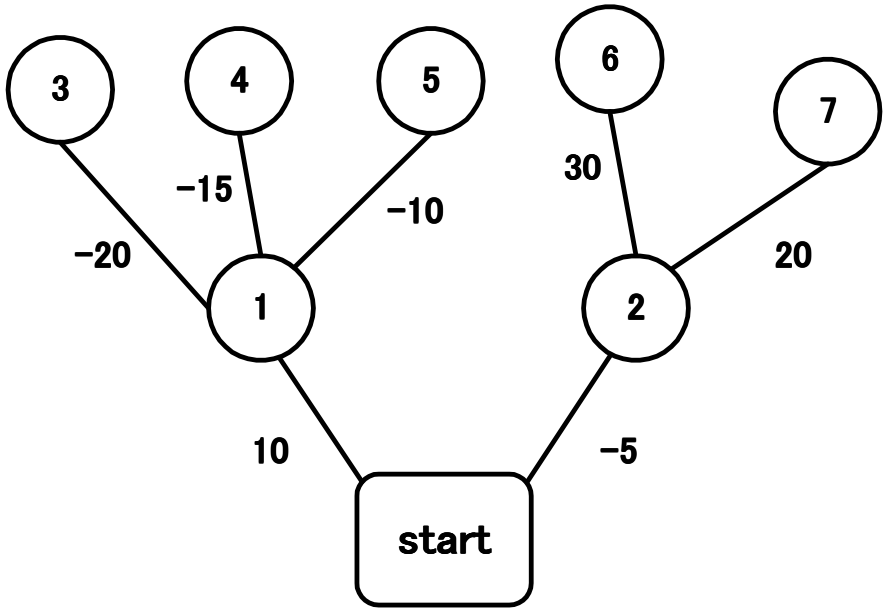


Fig.4-2　Expected return

うに即時報酬を最大化するように行動することは，将来の報酬を得にくくし，実際には収益が減少することになる．言い換えるとγが 1 に近づくに従い，将来の報酬をさらに考慮に入れることが目的となり，エージェントはより長期的な展望を持つようになる．

4･5･4　マルコフ性

前項では強化学習の目的が，エージェントがおかれている状態s_tにおいて，将来に渡ってトータルで多くの報酬が得られる行動a_tを探し出す，言い換えると収益を最大にする行動を探し出すことであることを述べた．

それではここまで述べてきた『状態』とはどのようなものを意味するのだろうか？強化学習における状態とは，どのようなものであれ，エージェントが利用可能な情報全てのことを言う．

例えばセンサから出力された電圧量をもって現在の状態と言ってもいいし，ロボットが右に向いている様子を一つの状態，あるいは気分が良いのか悪いのか，ということを一つの状態として取り扱ってもいい．

しかし強化学習を適用する上で，この情報には一体どのようなものが含まれていることが望ましいのだろうか？例えば，砲弾が飛んでいる風景のある一瞬を捕らえたとする．これに対して外から観察した情報として，その位置情報を取得しただけでは，砲弾がどちらの方向に飛んでいるかはわからない．つまり砲弾の情報としては，その位置および速度を一つの組として考えることで，完全な情報となりうる．

理想的な情報とは，状態信号として，過去の情報をコンパクトに集約し，さらに過去の関連情報を全て保持していることが望ましい．砲弾の例で言えば速度情報があることで，どちらから飛んできたのかという過去の情報を含ませることができる．

このように全ての関連情報をうまく保持することのできる状態信号は，マルコフ的である，あ

るいはマルコフ性を持つという．このような信号の例としては，例えばオセロの局面の情報などがそうである．過去にどのような経緯で現在の局面になったにせよ，現在の局面に全ての情報が集約されており，現在の局面のみで次に打つ手を決定することができる．（厳密に言えば過去の手を見て相手の癖がわかるとすれば，癖という情報が欠落しているが）

強化学習では，多くの場合このようなマルコフ性を持つ状態信号を仮定して定式化する．ここで実際にマルコフ性の形式的定義を行う．数学的表現を簡単にするため状態の数と報酬の大きさを有限とし,時刻tにおいて行動をとったとき,次の時刻t+1の状態がどのようになるかを考える．マルコフ的でない一般の場合には，過去の履歴すべてを考慮しなければならず，時刻 t+1 における状態s'で報酬rを得る確率を数式で表現すると，次のようになる．

$$P_r\{s_{t+1}=s',r_{t+1}=r\,|\,s_t,a_t,r_t,s_{t-1},a_{t-1},r_{t-1},\cdots,r_1,s_0,a_0\} \tag{4-5}$$

一方，状態信号がマルコフ性を持つなら，t+1 における環境の応答は t における状態と行動表現のみに依存することになり，このときには上式は

$$P_r\{s_{t+1}=s',r_{t+1}=r\,|\,s_t,a_t\} \tag{4-6}$$

となる．つまり，全てのs',rと履歴$s_t,a_t,r_t,\cdots,r_1,s_0,a_0$に対し，(4-6)が(4-5)に等しいとき，そしてそのときに限り状態信号がマルコフ性を持ち，マルコフ状態であると言われる．

もし環境がマルコフ性を持ち，現在の状態と行動が与えられるなら，その 1 ステップダイナミックス(4-6)から次の状態と行動を予測することが出来る．そしてこの式の反復計算を行うことにより，現時点までの完全な履歴が与えられた場合と同様に，現在の状態のみから将来の状態と期待される報酬の全てを予測することが出来る．このことから，マルコフ状態であることは行動を選択するために十分な情報を持ち合わせており，強化学習に際して非常に重要な要素である．

仮に状態信号が非マルコフであっても，部分的にマルコフであると仮定して学習を行うことが多くその有効性も確かめられている [19].

強化学習でこのマルコフ性が重要視される理由は，意思決定と価値が現在の状態のみに依存した関数であると仮定されているからである．

4･5･5　マルコフ決定過程

マルコフ性を満たす強化学習タスクはマルコフ決定過程(MDP)と呼ばれる．そして状態と行動の空間が有限であるならば，有限マルコフ決定過程(有限 MDP)と呼ばれる．有限 MDP は強化学習理論において非常に重要である．有限 MDP は状態と行動の集合と，環境の 1 ステップダイナミクスから定義される．任意の状態と行動，s と a が与えられたとして，次の状態がs'である確率は

$$P^a_{ss'}=P_r\{s_{t+1}=s'\,|\,s_t=s,a_t=a\} \tag{4-7}$$

であり，これらの量は遷移確率と呼ばれている．同様にして，現在の任意の状態 s と行動 a が与えられたとして，次の報酬の期待値は

$$R^a_{ss'}=E\{r_{t+1}\,|\,s_t=s,a_t=a,s_{t+1}=s'\} \tag{4-8}$$

である．本研究ではマルコフ性を満たす問題を扱う．

4･5･6　価値関数

ここまで，強化学習では将来的に得られる累積的な報酬を最大化することを目的として学習を行うことを述べ，その中でほとんど全ての強化学習理論がマルコフ過程において成り立っていることを述べた．

ここでは実際に累積的な報酬を最大にするにはどうすればいいのかを，定式化する．強化学習のほぼすべてのアルゴリズムでは価値関数に基づく評価を行っている．価値関数には大きく分けて，『状態価値関数』と『行動価値関数』の二つがあるのだが，順に説明を行う．

まず状態価値関数とは，エージェントがある状態にいることがどれだけ良いのか，というのを評価する関数である．ここで言う「どれだけよいか」という概念は，現在の状態にいた場合，その後期待される収益はどの程度なのか？ということを表している．もちろん，エージェントが受け取ることを期待できる報酬は，エージェントがどのような行動を取るかに依存する．したがって，価値関数は特定の『方策』に関して定義される．方策については 3･4･2 項で示したが，もう一度示すと，状態 s で行動 a を取る確率 $\pi(s,a)$ への写像であった．つまりある状態 s で方策にしたがった場合に，受け取れる期待収益を状態価値関数 $V^{\pi}(s)$ と表し次式で定義される．

$$V^{\pi}(s)=E_{\pi}\{R_t \mid s_t=s\}=E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t=s\right\} \tag{4-9}$$

ここに E_{π} は方策 π の元で得られる期待値を意味する．

同様にして行動価値関数について示すと，行動価値関数とはある状態においてある行動を取ることがどれだけ良いか，というのを評価する関数である．

そこで同様に，方策 π の元で状態 s において行動 a を取ることの価値を行動価値関数 $Q^{\pi}(s,a)$ で表し，状態 s で行動 a を取り，その方策 π に従った期待報酬として定義する．

$$Q^{\pi}(s,a)=E_{\pi}\{R_t \mid s_t=s,a_t=a\}=E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t=s,a_t=a\right\} \tag{4-10}$$

4･5･7　最適価値関数

ここまで述べてきたように，強化学習問題のタスクを解くことは，将来的に得る累積的な報酬を大きくするような方策を見つけることを意味する．この方策について有限 MDP に対しては，以下のようにして最適方策を正確に定義することができる．全ての状態に対して，方策 π の期待収益が π' よりも良いか同じであるなら，π は π' よりも良いか，同じであると定義される．言い換えるならば，全ての $s\in S$ に対して，$V^{\pi}(s)\geq V^{\pi'}(s)$ であるなら，そのときに限り，$\pi\geq\pi'$ である．他の方策よりも良いか，それに等しい方策が常に少なくとも 1 つ存在し，これが 1 つの最適方策である．最適方策は 1 つ以上存在するかもしれないが，全ての最適方策を π^* と記す．最適方策群は，最適状態価値関数と呼ばれる同じ状態価値関数を共有する．最適状態価値関数は，すべての $s\in S$ に対して

$$V^*(s)=\max_{\pi} V^{\pi}(s) \tag{4-11}$$

と定義される．これらは行動価値関数についても同様のことが言えて，最適行動価値関数Q^*は，すべての$s \in S$と$a \in A$に対して

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{4-12}$$

と定義される．この関数は，状態行動対(s,a)に対して，状態sにおいて行動aを取り，その後に最適方策に従うことに対する期待収益を与える．よってV^*を用いて，Q^*を次のように書くことができる．

$$Q^*(s,a) = E_{\pi}\{r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a\} \tag{4-13}$$

4･6　TD 学習

本章では価値関数を求める手法として代表的な手法である TD 学習について述べる．また TD 学習の中で広く使われている学習アルゴリズムのうち Q-Learning および Actor-Critic アルゴリズムについて説明する．

4･6･1　TD 学習とは

TD 学習とは経験強化型の学習であり，環境との相互作用を通じて価値関数を推定する．先に述べたように状態価値とは，その状態以降，期待される収益はどの程度なのか，つまりその後の報酬和で表わすことができた．つまりある時刻の価値は

$$V(s_t) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{4-14}$$

で表わせる．一方，次の時刻の予測値は

$$V(s_{t+1}) = \quad r_{t+2} + \quad \gamma r_{t+3} + \gamma^2 r_{t+4} + \cdots \tag{4-15}$$

である．現時点と次の時刻の予測値の間に

$$V(s_t) = \quad r_{t+1} + \gamma V(s_{t+1}) \tag{4-16}$$

の関係があり，この誤差量Δ

$$\Delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{4-17}$$

を考える．これを TD 誤差といい，この誤差を 0 に近づけていくという方法で学習を進める．上式で示したように TD 誤差とは，現在の状態の評価値と実際に行動してみて得られた価値を比較して，その状態の評価が正しかったかどうかという誤差である．

この TD 誤差が正の時には，見積もっていたよりも自分がいる状態はよかったということであり，負の時には見積もりよりも悪かったということになる．TD 誤差を求めて，現在の評価値を更新するが，この更新式は

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \tag{4-18}$$

という式になる．ここで現れる α は学習率と呼ばれる学習の早さを指定する 0≦α≦1 の係数である．TD 学習は，環境のモデルがわからないため，エージェントは実際に環境と相互作用してみる必要がある．時刻 t における環境との 1 回の相互作用から得られる経験を用いて，現在の状態の

価値を更新する．

以上の TD 学習の様子を例を挙げて説明する．

今 Fig.4-3 に示すような状態が遷移する環境を考える．エージェントは S0 をスタートとし，左右に動きながら最短でゴールに到達する方策を探し出すことを目的とする．エージェントはゴールへ到達した時点で環境から報酬を得る．ゴールへ到達した時点でエージェントはスタートに戻るエピソードタスクとする．

はじめはそれぞれの状態の評価値は定まっていないので全ての状態の評価値を同一の値で初期化しておく．最初の試行(エピソード)で，エージェントが S4 に到達し，その後右に行動をしたとする．すると式(4-18)に基づき，状態 4 の価値$V(S4)$の値が大きくなる．

次の試行で，エージェントが S3 に到達し，右に行動し S4 に到達したとすると，その際には報酬を得られないが，式(4-18)により価値を求めると

$$\begin{aligned} V(S3) &\leftarrow V(S3)+\alpha\left[r_{t+1}+\gamma V(S4)-V(S3)\right] \\ &= 0+\alpha\left[0+\gamma V(S4)-0\right]=\alpha\cdot\gamma V(S4) \end{aligned}$$

となり$V(S4)$が先の試行である正の値を持っているのでそれに引きずられる形で状態 3 の価値$V(S3)$の値が大きくなる．

このようにして価値が伝播していき，最終的に価値の高い方へ行動を選択していくことで収益を最大化させることができる．

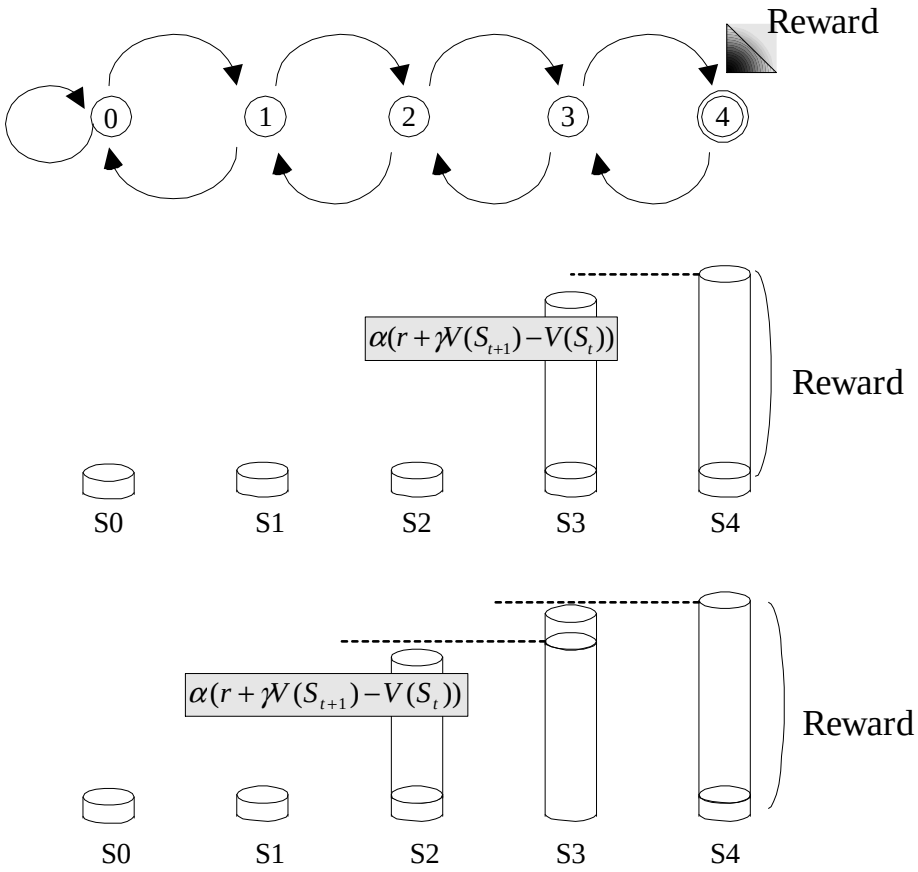


Fig.4-3　TD Learning

4･6･2　TD 学習の利点

TD 学習の利点としては，TD が環境のモデル，つまり報酬と次の状態の確率分布を必要としない点である．これは本論文では述べていないが，動的計画法(DP)に対する明らかな利点である．また，モンテカルロ法と比べると，モンテカルロ法がエピソードが終わらないと収益がわからないため，エピソードが終わるまで待つ必要があるのに対し，TD では 1 ステップ待つだけでよい．このことは非常に大きな利点となる．例えば非常に長いエピソードを待つタスクでは，エピソードの終わりまで待つと，学習があまりにも遅すぎ問題があり，またタスクが連続タスクの場合，エピソード自体が存在しないからである．動的計画法やモンテカルロ法については本論文では述べていないが，これらは TD 学習が他にはない利点である．

さらに TD アルゴリズムでは学習の収束性が保証されている点が大きな利点である．TD アルゴリズムでは，学習率が十分に小さい定数ならば，V^{π} へ収束すること，そして学習率が時間とともに減少するような場合，確率 1 で V^{π} へ収束することが証明されている．

4･6･3　Q-Learning

TD 学習をベースにそれらを拡張した強化学習アルゴリズムとして，Q-Learning がある．

Q-Learning は方策オフ型の TD 制御であり，状態行動対を一つのセットとして評価する．評価方法がわかりやすく，実装が容易な点から広く用いられている．

Q-Learning の特徴として状態と行動を一つのセットにして評価する点があげられるのだが，これらを例を挙げて説明する．

今 Fig.4-4 に示す迷路問題を考える．

状態 S0 をスタートしエージェントが上下左右に動きながら，ゴールまでの最短経路を Q-Learning により発見することを考える．

ここで報酬を以下のように設定する．

1. ある行動を起こして S15 に到達したら+10 の報酬
2. それ以外は－0.1 の報酬（無駄な動きであると考える）

Q-Learning では報酬を手がかりに行動価値関数 $Q(s_t, a_t)$ という値を求めていく．この行動価値関数は先にも述べたように，ある状態 s_t で行動 a_t を取ったときの良さというものを示している．例えば S14 では左右に動ける．これは左に動く行動のよさを $Q(S14, 左)$，右に動く行動のよさを $Q(S14, 右)$ とすると，S14 で右に行くとゴールに到達して報酬を得られることから，学習をしていくと $Q(S14, 右)$ が非常に大きくなり，相対的に $Q(S14, 左)$ が小さくなる．このように Q-Leanring では全ての状態行動対について行動価値関数を求めていく．この行動価値関数を実際に求めるには TD 学習を状態行動対に拡張した以下の式により求める．

$$Q(S_t, a_t) \leftarrow Q(S_t, a_t) + \alpha\left[r_t + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, a_t)\right] \tag{4-19}$$

式(4-18)にも示したように，TD 学習では遷移した次の状態の評価値をみるが，Q-Learning ではその状態について複数の行動に関連づけられた評価値を持つため，その中で最大のもの $\max_a Q(S_{t+1}, a)$ をみる．こうすることで，遷移した先の情報を含む価値というものを手に入れていくのである．

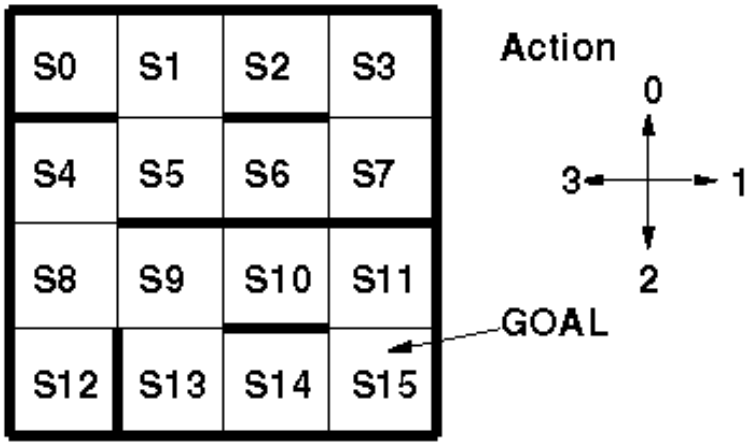


Fig.4-4　Maze problem

行動価値の更新の方法は以上であるが，その値に応じてどの行動を選択するかということが問題となる．行動選択の代表的な手法として ε-greedy 選択と Boltzman 選択が提唱されている．

ε-greedy 選択・・・定められた確率 ε でランダムに行動し，(1-ε)の確率で大きな方の Q 値を持つ行動を選択をする．

Boltzman 選択・・・状態におけるそれぞれの行動評価値に応じて，$\exp(Q(s,a)/T)$に比例する確率で行動を選択する．

以上のように Q-Learning の最大の特徴は，状態と行動をセットにして評価を行っていることであり，これによって，行動に対して直接的に評価を行うことができる．しかし，その状態においてとることのできる行動の数が多くなると，用意しなければならないセットの数が非常に膨大になってしまうため，連続的な行動出力を要求されるような環境に対しては，Q-Learning よりも次項で示す Actor-Critic 法が一般的に用いられる．

4･6･4　Actor-Critic

Actor-Critic アルゴリズムは TD 学習を用いた最も初期の強化学習システムで用いられていたものである．Actor-Critic アルゴリズムでは Fig.4-5 に示すように，評価を行う部分(Critic)と行動選択を行う部分(Actor)を完全に独立にさせて考えることから先に挙げた Q-Learning などにはない以下の利点がある．

行動選択に最小限の計算量しか必要としない．例えば連続値行動のように，可能な行動の個数が無限大である場合，Q-Learning などのような行動価値を学習するアルゴリズムでは 1 つの行動を選び出すために無限集合のなかを探索することになってしまう

確率的な行動選択を学習することが出来る．すなわち，いろいろな行動に対して，それを選択するような最適確率を学習することが出来る．

このようなことがなぜ可能なのかについて説明する．今 Fig.4-6 に示すような状態が遷移する環境を考える．これは様々な行動の中のある一部だと考えると，連続値を行動出力として要求する環境というのは，このような行動数が無限個あるものと考えることができる．このような環境において，Q-Learning を適用させようとすると，(状態数×行動数)個の Q 値を格納させておかなければならない．それらを格納するメモリの問題だけでなく，学習時間も膨大にかかる．これに対して，Actor-Critic では，評価を行う部分と，行動選択の部分を完全に独立させて考える．

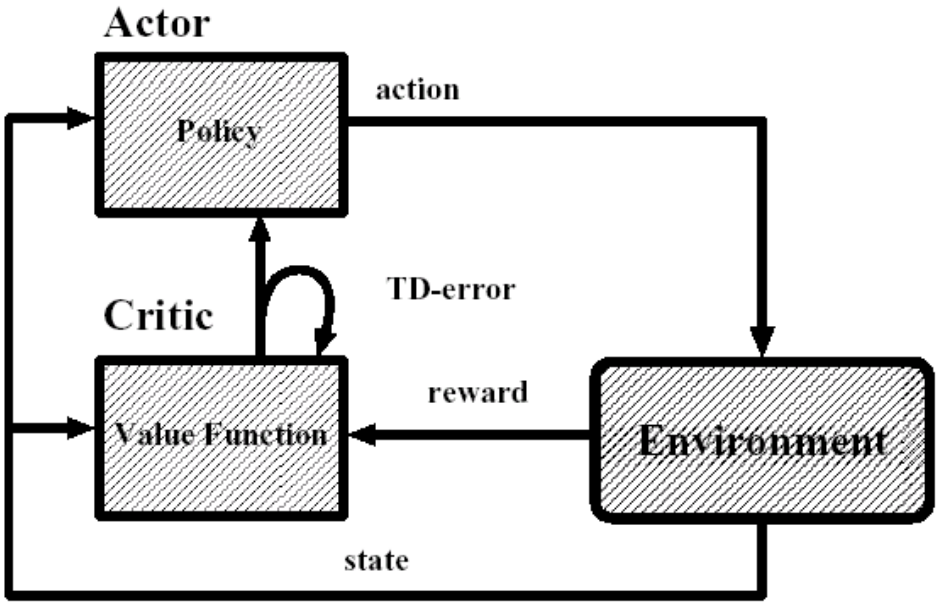


Fig.4-5 Actor-Critic

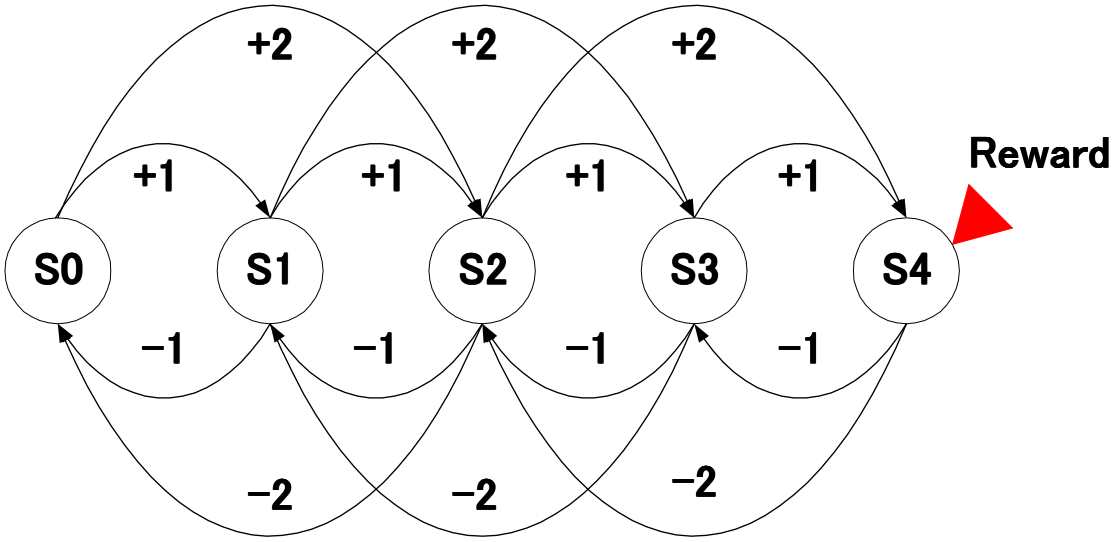


Fig.4-6 Actor-Critic examination

例えば，学習における最初の試行で，状態が S0→S1→S2→S4 と遷移して報酬を得たとする．すると状態の価値は TD 学習で示した式(4-18)により求めると，Fig.4-7 のように S2 の評価値が上がる．ここで S2 において，行動+2 を選択する確率を上げる．actor 部は，確率を設定できるものであればどのようなものでもよいが，正規乱数を用いているものとする．初期値として，中心値 0，標準偏差 1 という値があったとする．正規乱数は Fig.4-8(1)に示すような頻度で発生する．赤い点は，正規乱数を発生させて行動を選択させた際に，S2 で発生した乱数であると考える．

このような乱数が発生し，状態が予期していたよりもよかったと判明した．この状態評価は TD 誤差を用いる．この行動を選択したことで，予期していたよりも状態がよかったと判明したの

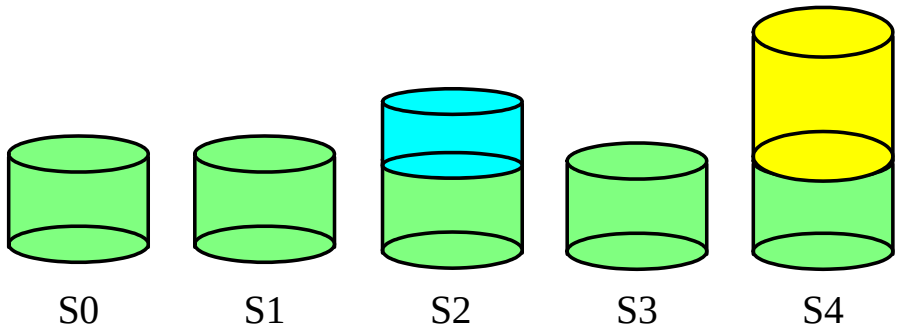
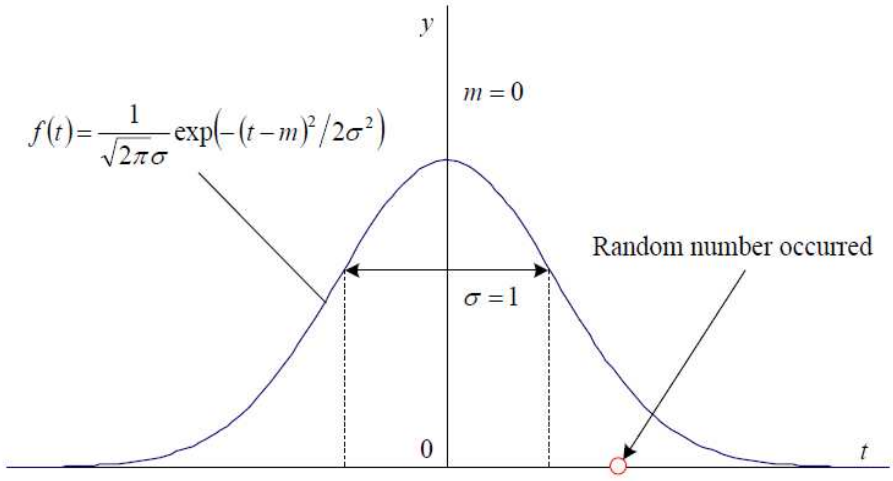


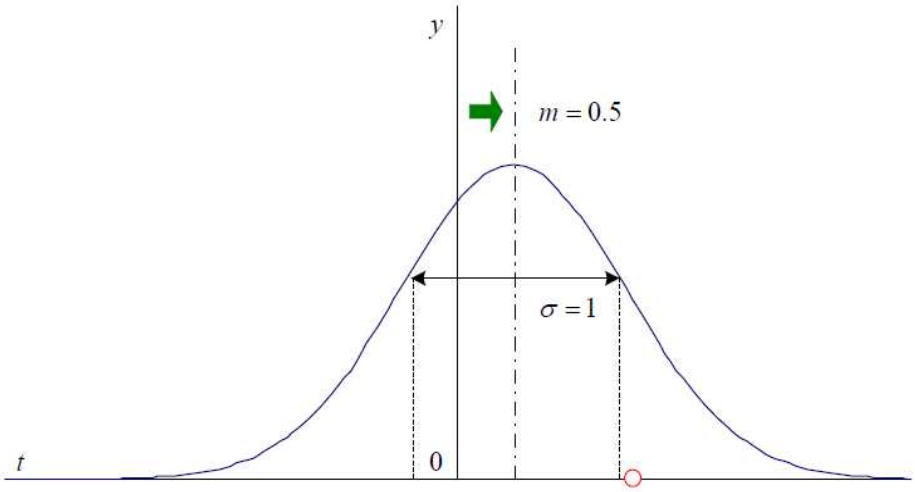
Fig.4-7 State value function

で，+2 という行動を選択する確率が大きくなるようにする．具体的には正規乱数の発生する際の中心値を+2 の方向に寄せ(Fig.4-8 (2))，発生した乱数が標準偏差よりも内側だったので，標準偏差の値を大きくする(Fig.4-7 (3))．こうすることによって，+2 が選択される確率を高くすることができる．

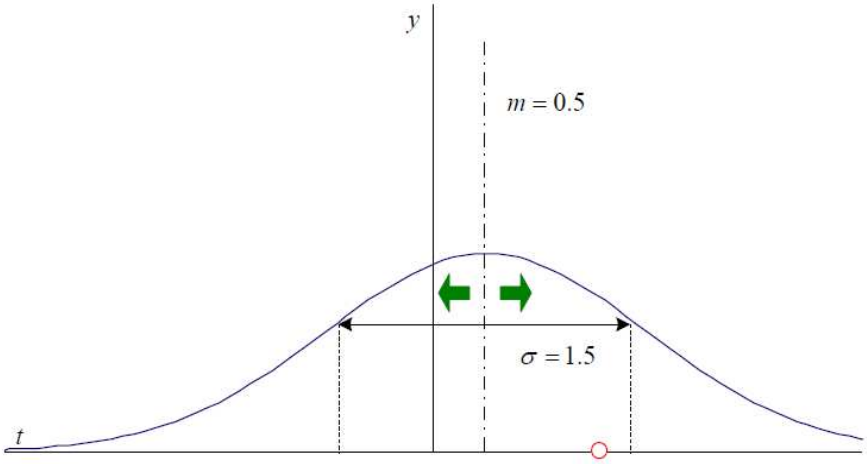
今回は，標準偏差の外側であったので，標準偏差の値を大きくしたが，TD 誤差が正であり(すなわち，見積もりよりも状態がよかったと判断され)，かつ発生した乱数が標準偏差の内側であった場合は，標準偏差の値を小さくして，発生する乱数の幅を狭める．このような正規乱数を用いる利点は，今回のような離散的な行動を用いる際にはあまり明確にはならないが，連続的な値を行動として出力するときには有効である．なぜなら Q-Learning のような行動に対する評価値をそれぞれ定める際には(状態数×行動数)個の値を格納しておく必要があったのに対して，Actor-Critic で正規乱数を用いた際に必要となるのは，(状態数×2(中心値と標準偏差))だけだからである．



(1)



(2)



(3)

Fig.4-8 Normal random number

4-7　実験システム

本節では，本研究で用いた実験システム（実験機，シミュレータ）についてその概要を述べる．

4-7-1　システム構成

Fig.4-9 に示すように，本研究の実験に用いるシステムは，CPU・モータドライバ・モータ・減速機が組み込まれたモータユニット（AI Motor-601，Megarobotics 社）（以下 AI Motor）4 個，PSD センサを用いたポジションセンサシステム（浜松ホトニクス）1 台，パーソナルコンピュータ（Face 製オリジナルコンピュータ）1 台と各種 I/O ボードから構成されている．

個々については後に述べるが，Fig.4-10 に示すようにモータユニット 4 つを組み合わせ，多関節の芋虫型ロボットを作成した．各モータユニットには PSD センサが取り付けられており，電圧データをロボット側面からカメラにより捕捉することにより，各点の x-y 座標（2 次元座標）を測定することができることから，ロボットの移動距離を算出することができる．(Fig.4-11 参照)

カメラから取り込まれたデータは，ポジションセンサアンプで増幅され，A-D ボードを介してパーソナルコンピュータに送られる．

一方モータユニットへの指令は，RS-232C 通信によりパーソナルコンピュータから出力角度を指令する．

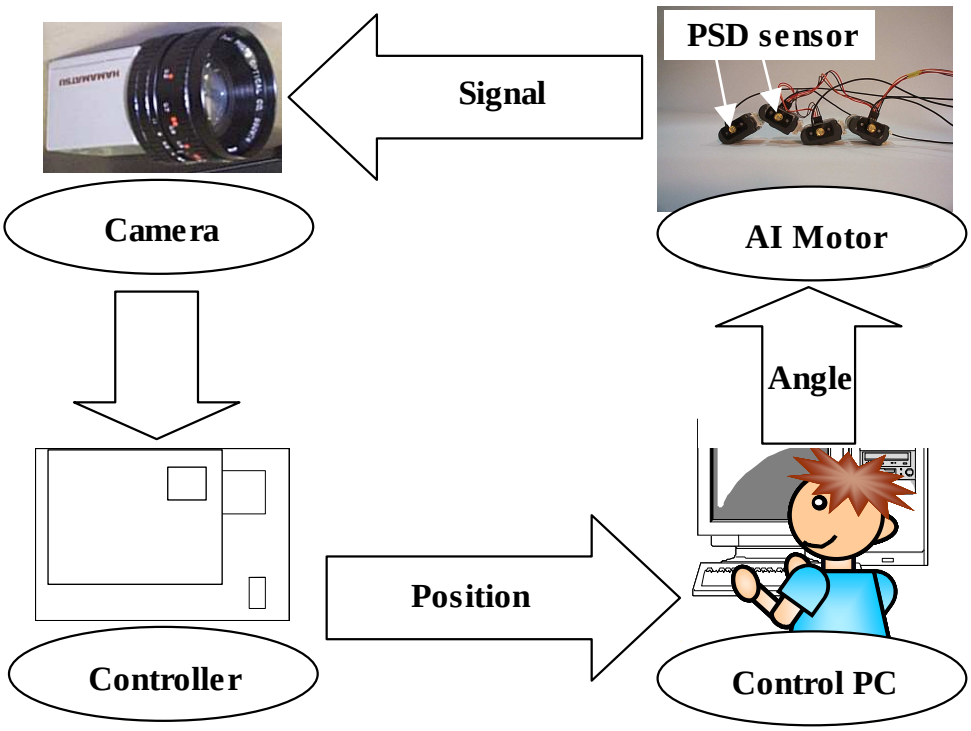


Fig.4-9　Experiment System

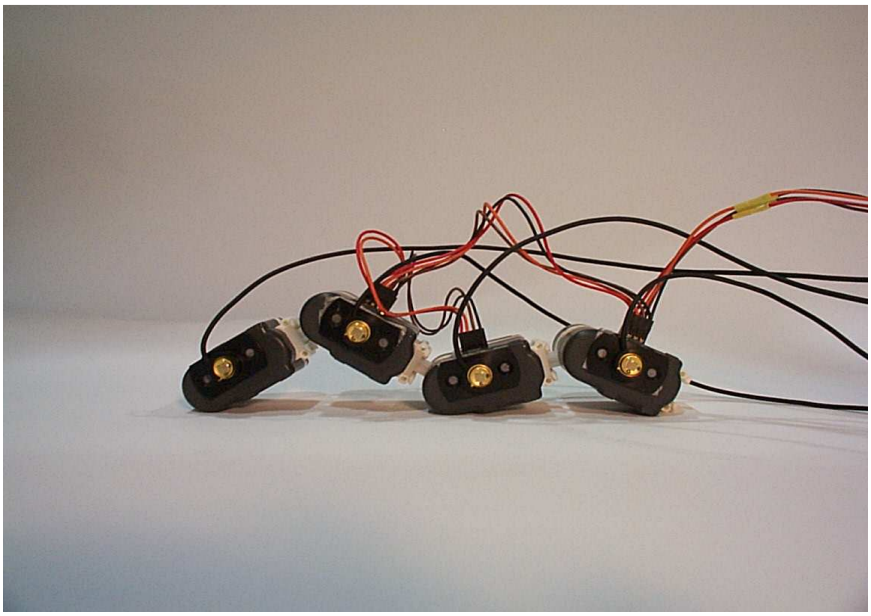


Fig.4-10 Experiment System

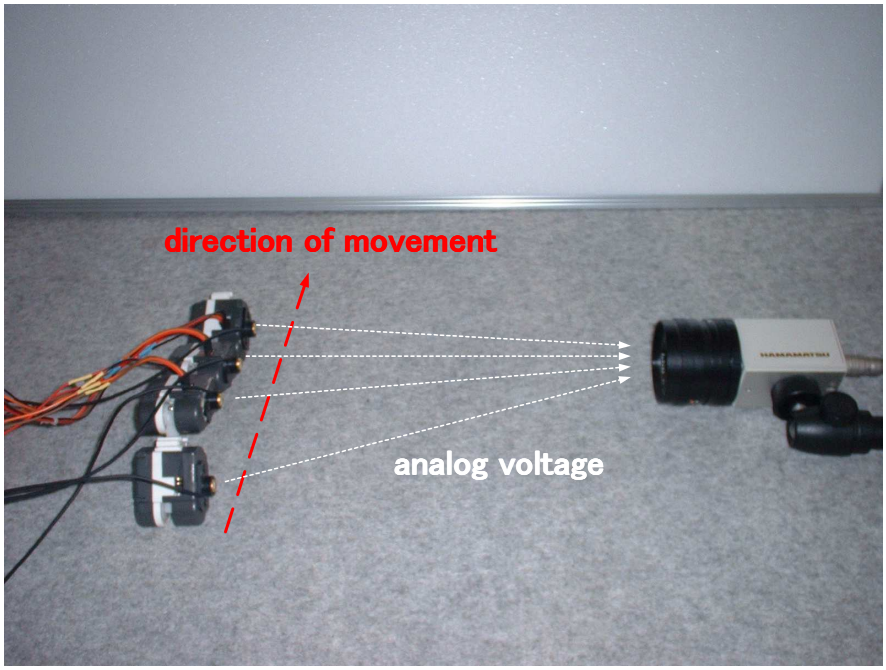


Fig.4-11 PSD Sensor

4･7･2　実験装置の仕様

本節では実験装置の個々の仕様について述べる．

(1) モータユニット

モータユニットには Fig.4-12 に示す Megarobotics 社製の AI Motor-601 を用いた．モータユニッ

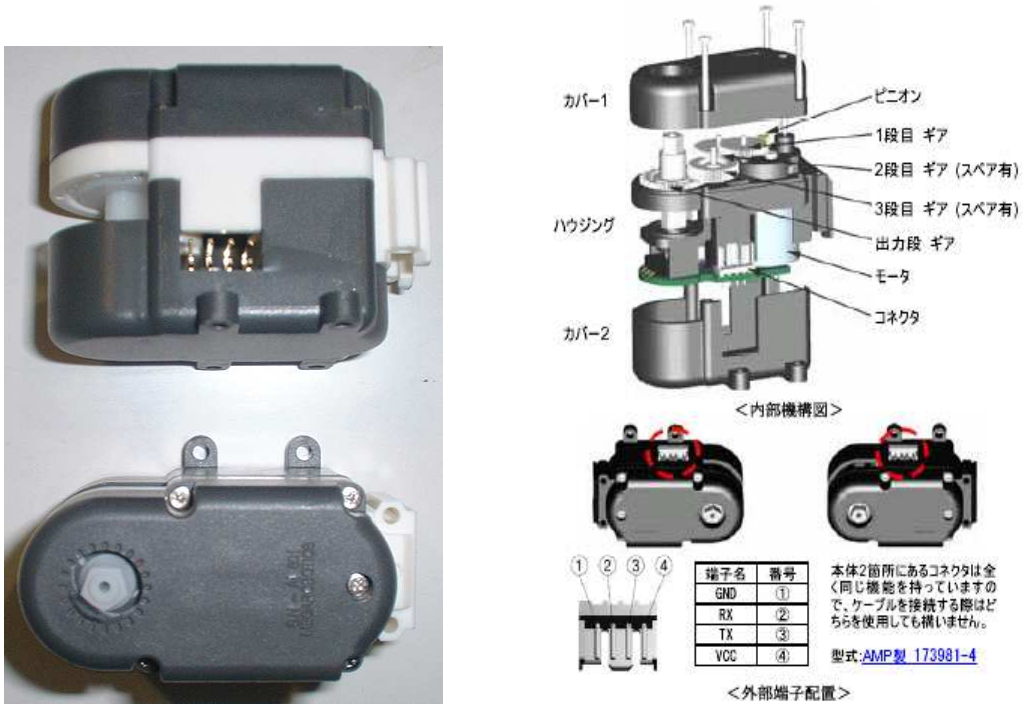


Fig.4-12　AI Motor-601

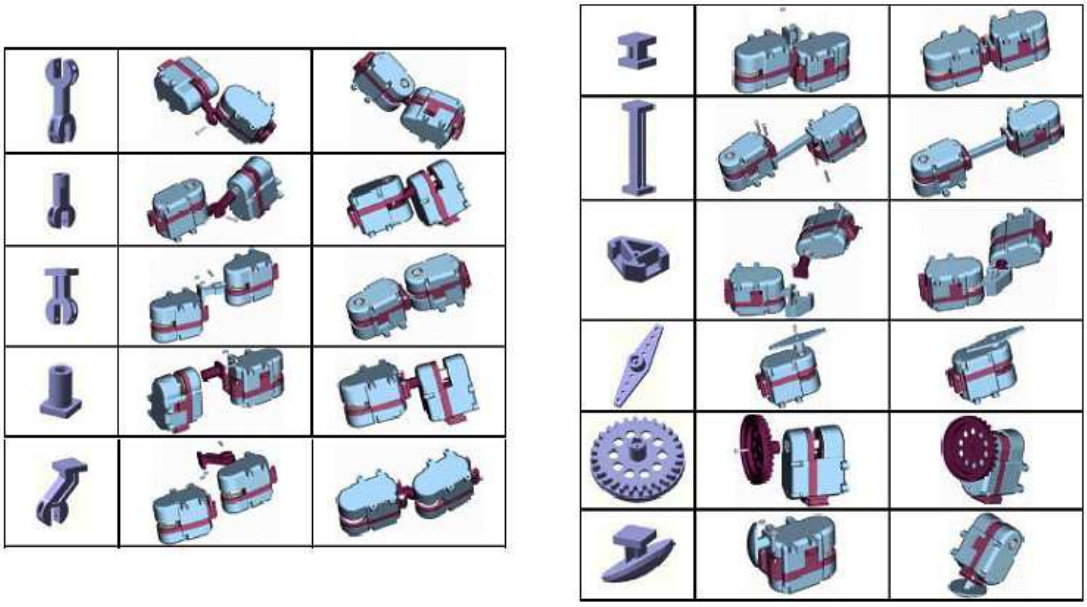


Fig.4-13　Joint parts

トには CPU・モータドライバ・モータ・減速機が組み込まれており，制御指令が UART（シリアル通信）で行われるため，シリアルポートを装備しているマイコンや PC から制御が可能なため，非常に扱いやすい．パッケージにはモータユニットの他，Fig.4-13 に示す各モータユニットを様々な形で連結できるようなジョイントパーツが含まれている．

以下にモータユニットの仕様を記載する．

主要定格及び仕様

機械的仕様	重量	40g
	サイズ	51.5×27.75×37.35mm
	接続ポイント	出力軸 2 箇所，ボディ 1 個所
	ギア	Ration:1/187 プラスチック製　ベアリングなし
	最大トルク	6kg・cm at 9.5V
	最大回転数	80rpm at 9.5V
電気的仕様	モータ	DC モータ　貴金属ブラシ使用
	最大電流	650mA at DC5V, 1000mA at DC10V
	UART 信号レベル	出力 High レベル：3.25～4.7V, Low レベル：0～0.6V 入力 High レベル：2.80～5.2V, Low レベル：-1.0～1.41V
	ロジック消費電力	10mA(4.5mA at Power Down Mode)
	電源電圧範囲	DC4.5～10.5V
	電流リミット	400mA～950mA
	位置分解能	Low Resolution：333.33deg/256=1.3deg High Resolution：166.67deg/256=0.65deg
通信方式	TYPE	UART
	通信速度	2400～460.8kbps
内部パラメータ	ID	0～30
	Baud rate	2400～460.8kbps
	Resolution	1.3/0.65deg
	Gain	P Gain, I Gain
	Over Current	400mA～950mA
動作モード	Position Send	255 段階の位置指令，5 段階の速度設定
	Position Read	出力トルク 0 で出力軸位置のモニタ
	Act Down	指定 ID の出力トルクを 0（停止）
	Wheel Act	16 段階の速度および正逆転指令で無限回転
	Power Down	全モータの省電力モードへの遷移

内部ブロック図

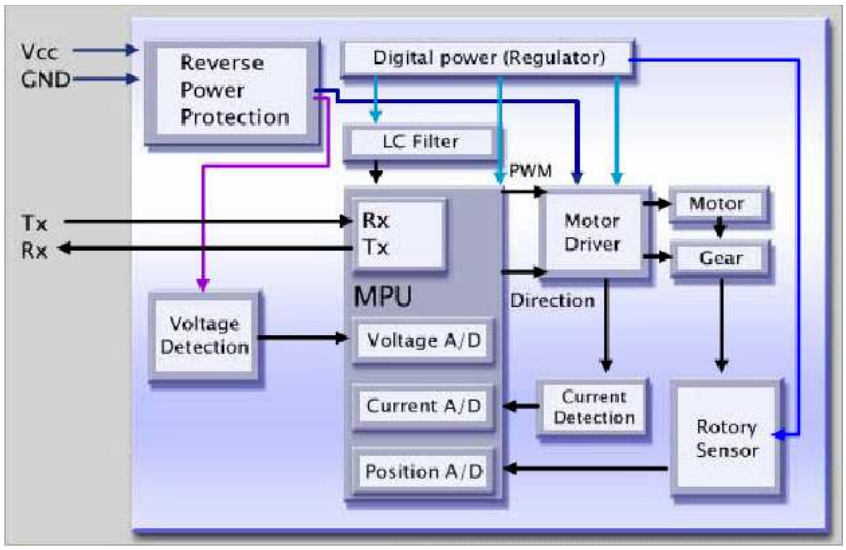


Fig.4-14 Block diagram

外部寸法

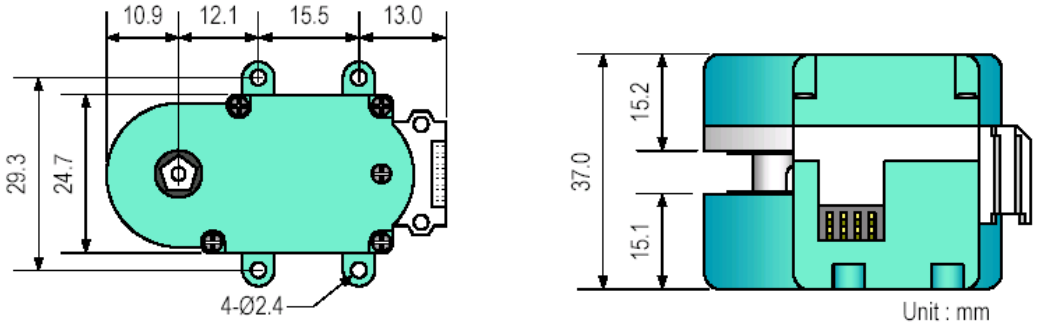


Fig.4-15 Outside dimension

内部機構図

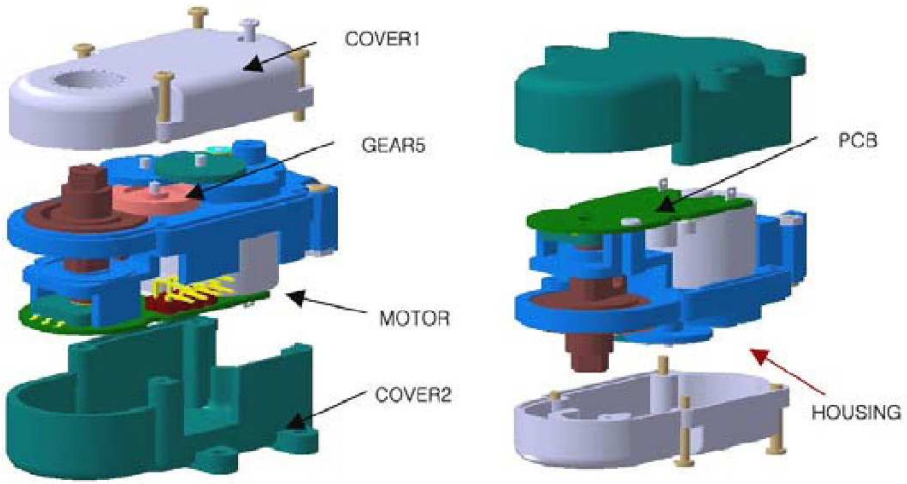
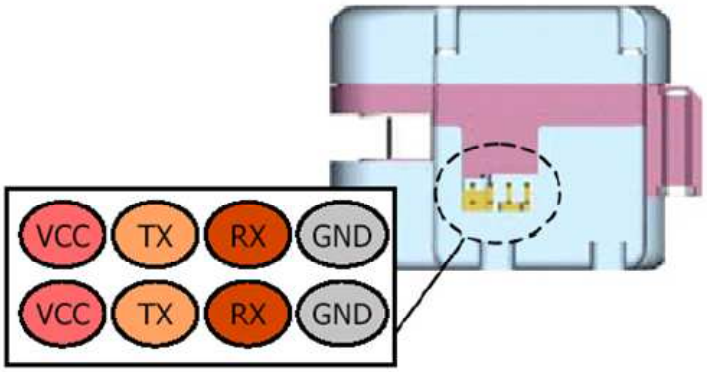


Fig.4-16 Internal mechanism

外部端子配置



Terminal	Color
VCC	Red
TX	Orange
RX	Brown
GND	Black

Fig.4-17 External terminal

動作モードとしては 360°の回転が可能な車輪モード，0～332.03°もしくは 0～166.07°いずれかのレンジにおける位置制御モード，位置や電流の変化をモニタできる位置取得モード，消費電力を最小化する省電力モードを持っている．また，内部パラメータ(ID・位置制御ゲイン・過電流リミッタ・通信速度)を変更できるため，ターゲットにあわせた微調整が可能である．

(2)AI Motor 用 PC I/F ボード

モータユニットの通信は TTL レベルで行われる．制御は PC の COM ポート経由で行うため，変換ボードを使用した．Fig.4-18 にボードの外形を，Fig.4-19 に回路図を示すが，このボードはモータへの電源を供給し，COM ポートの EIA232 レベルの信号を TTL レベルに変換して AI Motor の TX・RX 端子へ接続する．

また個々のモータには 0～30 の ID を持たせることができ，最大 31 個まで同一電源および UART 信上で並列接続することが可能である．さらにモータには内部で並列接続された 2 個のコネクタが装備されており，複数のモータを介して構成される配線をスマートに行うことができる．

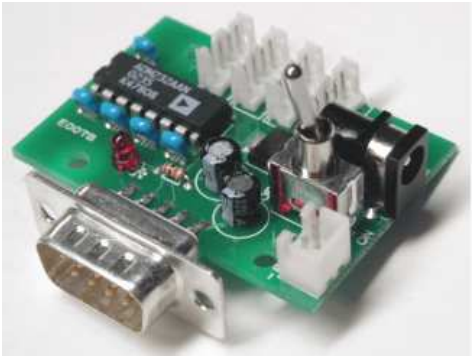


Fig.4-18 PC I/F board

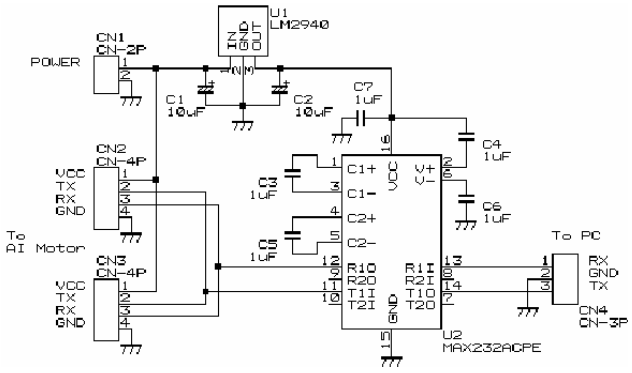


Fig.4-19 Circuit diagram of PC I/F board

(3)ポジションセンサシステム

各 AI Motor にはそれぞれ LED の発光装置が取り付けられており，それらが赤外線を発し，ロボット側部からカメラにて赤外線を検出する．これによって得られる位置データはポジションセンサアンプで増幅され，制御用コンピュータへ送られる．以下にポジションセンサとアンプの仕様を示す．

LED ドライバ仕様

LED ドライブ点数	最大 7 点
LED 点灯周波数	300Hz （146~300Hz 設定可能）
使用電源	内部：単三電池 ／1.5V x 4 本 外部：DC4.5V～7V
連続使用時間	約 20 分（LED7 点接続時） （Ni-Cd 電池使用時）
消費電流	1A（LED7 点接続時）

ポジションセンサ仕様

一般使用

検出器	半導体位置検出素子（S1880）
使用受光面寸法	10mm x 10mm
レンズマウント	C マウント
動作周囲温度	0℃～+40℃
保存周囲温度	-10℃～+50℃
動作，保存周囲湿度	90％以下（結露しないこと）
入力電源	100V
外形寸法及び重量センサヘッド	40(W) x 42(H) x 64(D)mm 約 140g
コントローラ	232(W) x 74(H) x 308(D)mm 約 3.4 kg

電気的使用

出力電圧：X 軸 Y 軸	-5V～+5V -5V～+5V
出力インピーダンス	500Ω±50Ω
外部クロック信号	TTL レベル
サンプリング周波数	内部モード：300Hz（標準）
推奨測定光量	光量レベル（Σ）＝4～8
位置検出誤差：ZONE A ZONE B	±1％ ±2％
光量変化による誤差	±1％（Σ8→4）
分解能	1／5000
ジッタ	±1／1000
ドリフト	±0.5%／DAY（ただし初期 30 分の変動を除く）

制御用ボード

A/D ボード

カメラセンサアンプからのアナログ信号をデジタル信号にしてパソコンに送る

（株）コンテック　　AD-12-16(PCI)

使用時の入力レンジ　：　±5V

分解能　　：　12bit

Fig.4-20 A/D board

制御用パーソナルコンピュータ

Face 製オリジナルコンピュータ

ハードウェア

CPU　　：Intel 製 Pentium4 2.53GHz 搭載

メモリ　　：512MB

ソフトウェア

OS　　：Microsoft Windows 2000

プログラミング環境　　：Microsoft Visual C++ 6.0

以上が 5 章で示す客観報酬に基づくロボットの前進行動獲得の実験装置である.

4･8　実験方法

本節では，PSD センサから得られる移動距離情報を報酬として用いてロボットに前進行動を獲得させる方法について述べる．本研究で用いた強化学習アルゴリズムは TD 学習をベースにした Q-Learning である．Q-Learning の理論については 4・6・3 で述べたが，状態行動対を一つの行動価値関数によって評価できるため，様々な問題に容易に用いることができ，その研究例も多い．また本研究のようなマルコフ性を満たすような環境の場合，その収束性が保証されていることから，各種比較検証を行う上で，より合理的に判断できると考えた．

学習手法には実機を用いロボットを動かしながらセンシング，学習，プランニング，行動を繰り返し行う On-Line 学習プロセスと，実機から得られた情報を蓄えておき，PC 上で学習を行う Off-Line 学習プロセスがあるが，本研究では，ロボットの疲労や学習時間の問題自体を考えることが本研究の目的ではなく，報酬と学習結果を多角的に検証するため，より多くの試行回数が必要である点から，学習は実機から得られた報酬情報を元にシミュレータを作成し，シミュレータを用いて Off-Line で学習を行い，その結果を実機に還元するという手法を用いた．

本章では，Q-Learning の適用方法，及びその学習プロセスについて，詳細を述べる．

4･8･1　Off-Line 学習プロセス

本研究では強化学習を用いてロボットに前進行動を獲得させる．前述のように，強化学習では何らかの行動に対して環境から報酬というスカラー量を受け取ることにより，一連の行動を獲得していく．本研究ではロボットがモータを動かすことにより，ある状態から何らかの状態に遷移した際に，その変化（移動距離）をセンサにより取得し，それらを報酬としてロボットに与えることで前進行動を獲得させる．

先にも述べたように，ロボットが何らかの動きをし，リアルタイムで環境から報酬を得て，それを元に学習を行い，学習結果から何らかの行動を起こす一連の学習過程をロボットを常に動かしながら行う On-Line での学習は，ロボットの疲労や学習時間を考えた場合，現実的ではない．そこで本研究では，実機から得られた報酬情報を元に簡易的なシミュレータを作成し，各種学習のシミュレーションを行い，比較検証を行った．その後，必要に応じて，実機にそれらの結果を還元し，実験ベースでの有効性も示した．以下にこの Off-Line 学習プロセスについてその手順の詳細を示す．

4･8･2　状態パターン

本研究では芋虫型ロボットシステムに強化学習を適用し，前進行動獲得を行う．Fig.4-21 に PSD センサを取り付けたロボットシステムの外形を示す．はじめに本ロボットの取りうる状態パターンについて述べる．本ロボットは 4 つのモータのうち，駆動する部分を Fig.4-22 に丸印で示す 2 ヶ所とし，Q-Learning のテーブル型の離散型データとして扱いやすくするために Fig.4-22 に示すように±90[deg]，±45[deg]，0[deg]の各々5 つの角度を指令値として用いた．この為ロボットが取りうる状態パターンは 25 通り(5×5)あることになる．Fig.4-23 にこれら 25 通りの全状態パターンを示す．本研究において便宜的に各パターンに 0～24 の状態番号を割り振った．

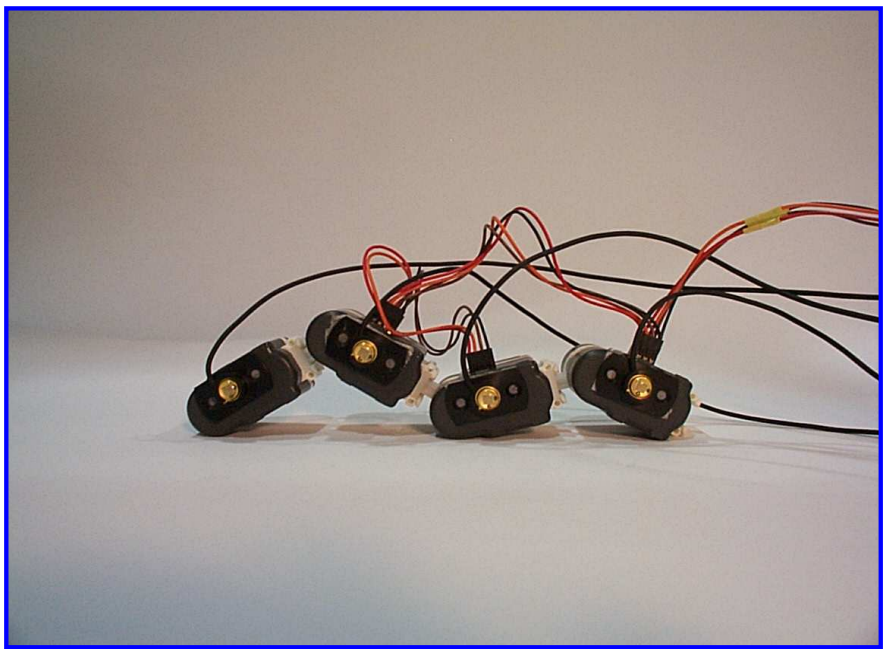


Fig.4-21 Experiment Systems

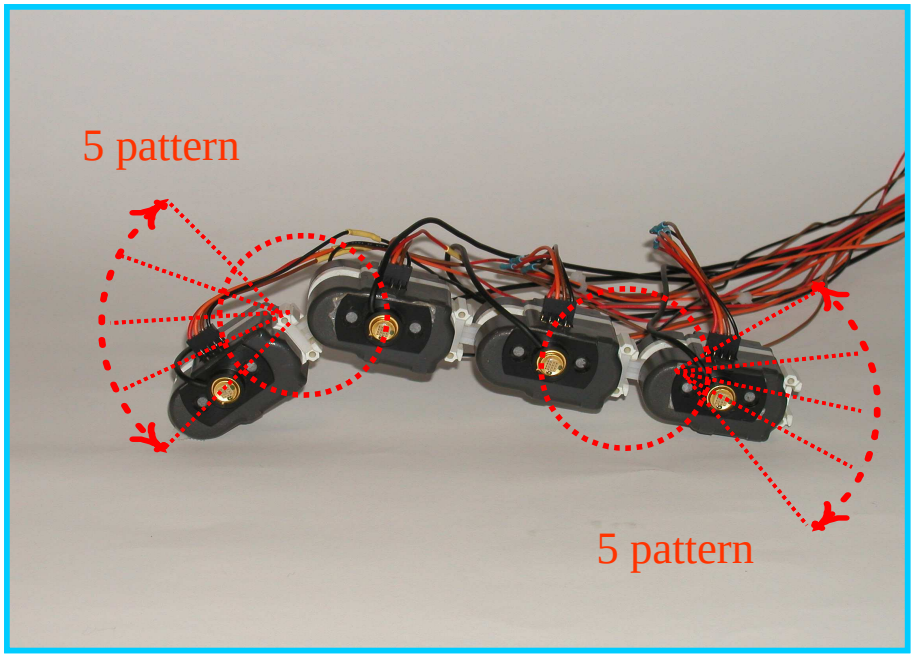


Fig.4-22 Driving method

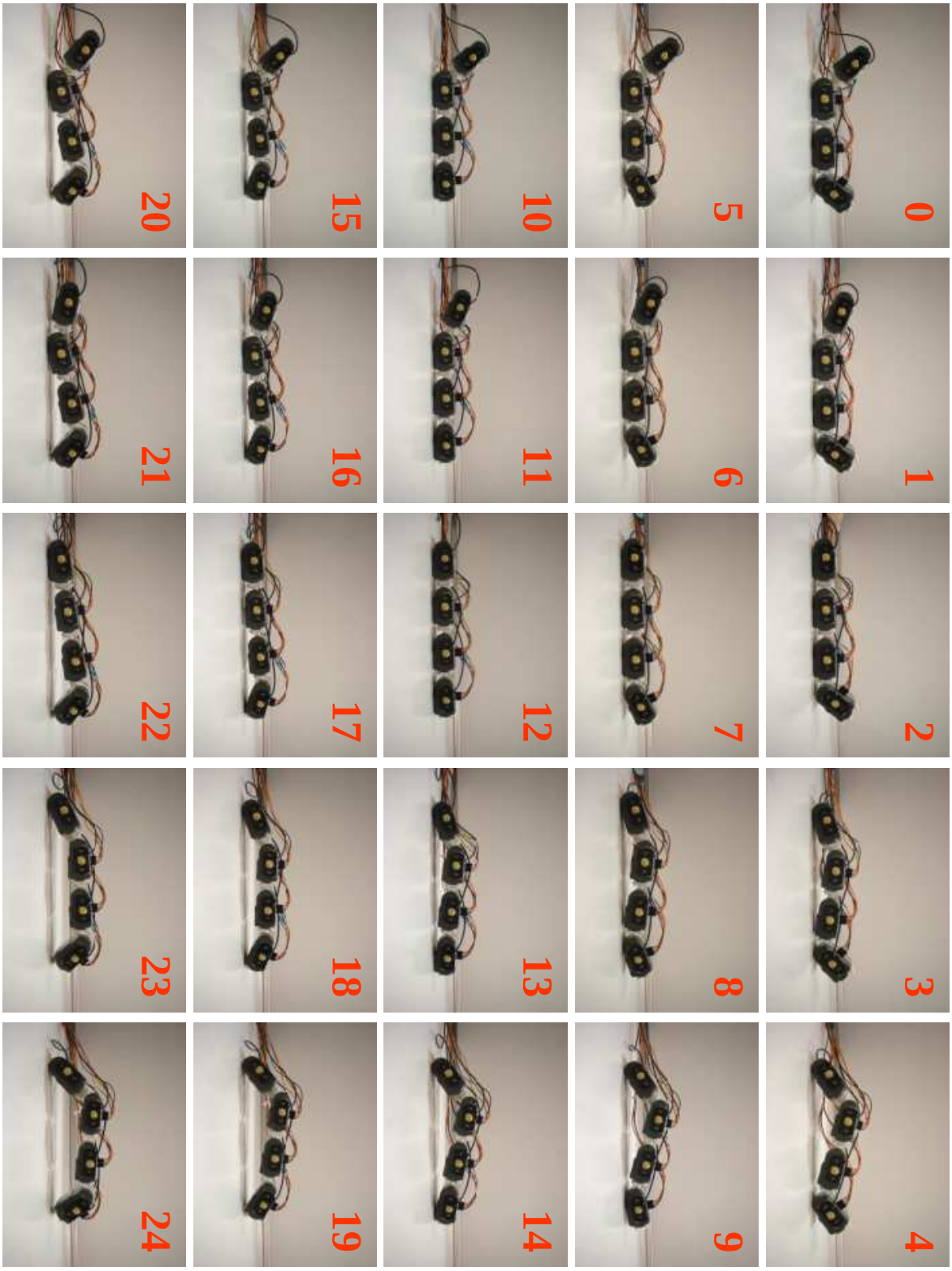


Fig.4-23 Robotic behavior

4･8･3　報酬の獲得

前項で示したようにロボットは 25 通りの状態パターンを持っている．ロボットがある状態 s_t から何らかの行動 a_t を起こし s_{t+1} に遷移した際に，ロボット側面に取り付けられた PSD センサにより移動距離を電圧データとして取得する．ここで電圧 V[v]と移動距離 L[mm]の間の関係は以下のようになっている．

L=○*V

ロボットに与えられる報酬 r_t は PSD センサの電圧値として学習を行う．Fig.4-24 に概念図を示す．なお本システムでは行動がそのまま遷移した先の状態となるため $a_t = s_{t+1}$ である．

ここで，便宜上報酬を二次元の配列として扱い，r [s_t][s_{t+1}]として表現する．例えば，Fig.4-3 に示したパターンの中で，状態 14 から状態 23 へ遷移した際に，センサから電圧として 0.289307 という値を得たとしよう．すると報酬の表現の仕方として

r [14][23] = 0.289307

というように表現できる．このように表現すると，全ての状態行動対について報酬情報を配列として格納できる．

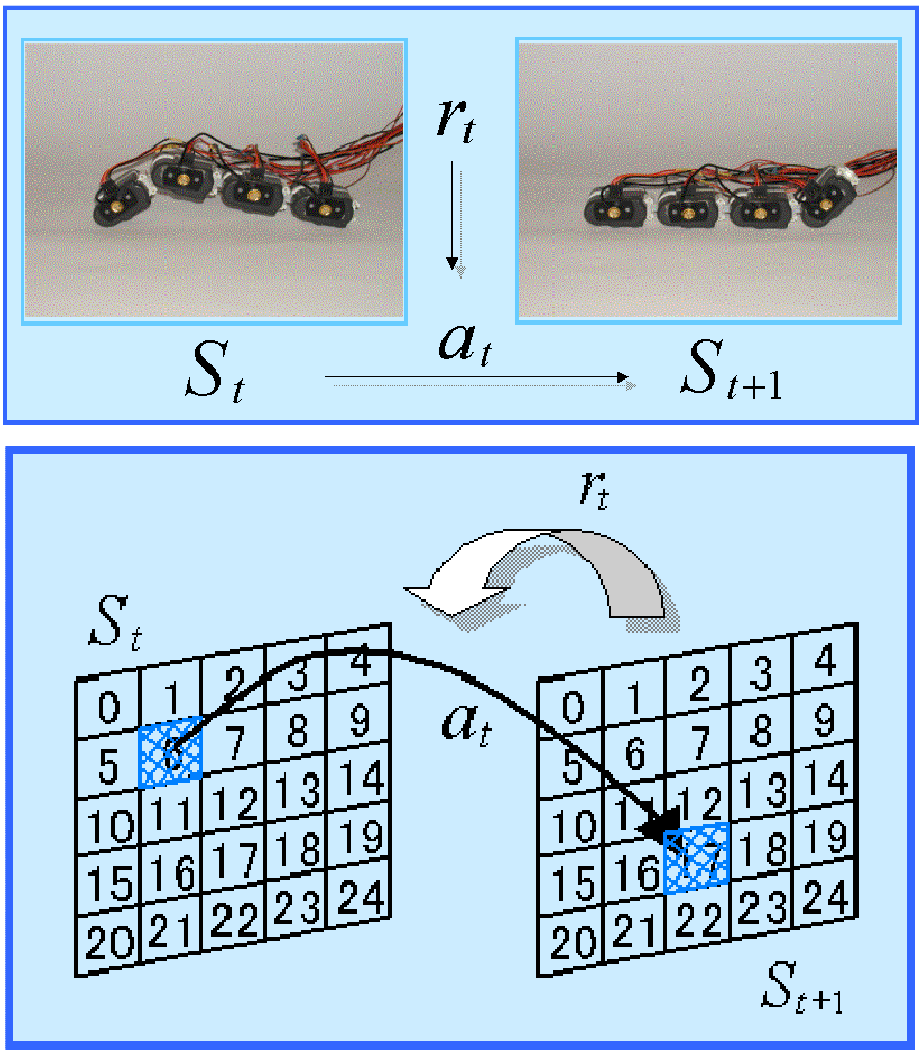


Fig.4-24 Conceptual diagram

r [0][0] = 0.001221
r [0][1] = -0.084839
r [0][2] = -0.123901
.
.
.
r [24][23] = -0.114136
r [24][24] = 0.003662

このようにして配列化された報酬情報は，ある 25 通りの状態から次の 25 通り分の情報があるため，全部で 625 通り(25×25)の情報があることになる．

実際に学習で用いるために，現実世界の摩擦などのランダム性を考慮して，625 通りの状態遷移パターンを 5 回ずつ実際にｚｘ動かし，それらを平均化し 625 通りの報酬情報を取得した．学習ではこれらのデータベース化された値を参照しながら学習を行う．

4･8･4　シミュレーション方法

上記で取得しておいた報酬情報を用いて Q-Learning を適用する．本項では具体的なシミュレーション方法について述べる．

はじめにロボットの初期姿勢を決めておく．本研究では Fig.4-23 からもわかるように 12 という状態がまっすぐの姿勢なので，状態 12 を初期姿勢と決めた．この状態からロボットにランダムに何らかの状態に遷移させる．例えばここでは 16 という状態に遷移したとする．すると先にデータベース化した報酬情報から

r [12][16] = 0.008545

という値を参照する．この報酬値を用いて Q-Learning において行動価値関数の値を計算する．式(4-19)で示したように行動価値関数$Q(s_t, a_t)$は

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

で表されるから，例えば学習率 α=0.9,割引率 γ=0.9 として実際に計算すると

$$Q(12,16) \leftarrow Q(12,16) + 0.9\left[0.008545 + 0.9 \max_a Q(12, a) - Q(12,16)\right]$$
$$= \quad 0 \quad + 0.9\left[0.008545 + 0.9 \cdot 0 - 0\right]$$
$$= \quad 0.0076905$$

となる．

このようにランダムな状態遷移→行動価値関数の計算の一組を学習回数 1 回として数える．

次に現在の状態は 16 にいるので，そこからまたランダムに状態を遷移させ，例えば 5 という状態に遷移したとする．すると同様に先にデータベース化した報酬情報から

r [16][5] = 0.125732

という値を参照し，同様に行動価値関数$Q(16,5)$という値を計算する．このようにして学習回数を重ねると共に選択した状態行動対の行動価値関数が求まっていく．

Q [0][0] = 0.002121
Q [0][1] = -0.024839
Q [0][2] = -0.154901
.
.
.
Q [24][23] = -0.214536
Q [24][24] = 0.003732

学習が何回か行われ行動価値関数がある程度求まったら，それらの行動価値関数をもとに行動を選択することになる．理想としては学習回数 1 回ごとに行動価値関数を元に何らかの行動をして評価することが望ましいが，実際にはメモリ量に限界があるので，学習回数 100 回おき，1000 回おき程度で行動価値関数を保存しておく．

行動選択の方法だが，状態 12 の初期姿勢をスタートとして，行動価値関数が最大の行動を 1,0000 ステップ行い，遷移させるたびに先の報酬データベースから報酬量を参照し足し合わせていく．つまりはじめ，

Q [12][0] = 0.03141
Q [12][1] = -0.030483
Q [12][2] = -0.204759
.
.
.
Q [12][23] = -0.235364
Q [12][24] = 0.002054

の中で，行動価値関数が最大のものが *Q* [12][4]であったとすれば，12→4 に状態を遷移させ，それに対応する報酬 *r* [12][4]を参照し，次に *Q* [4][0]～*Q* [4][24]の中から行動価値が最大の行動を選び仮に *Q* [4][18]が最大であれば 4→18 に状態を遷移させ，それに対応する報酬 *r* [4][18]を参照し先の報酬と足し合わせる．こうして 10000 ステップ間に受け取る報酬を平均化して，一歩当たりに進む距離というものを評価指標として用いる．学習が正常に行われれば，学習回数と共によりよい行動価値関数が求まっていくのでそれらを参照して，行動を行っていけば一歩当たりに進む距離が増加することになる．

4･8･5　実験方法

前項ではシミュレーション方法について述べたが，シミュレーションによって行動価値関数が最大の行動を選択していくことにより，状態が 12→6→2→・・・・・というように遷移していくと，最終的には 12→6→2→14→20→2→14→20 というように，ある決まったループを形成していくことになる．このループを本論文では便宜的に行動形態と呼ぶ．これらの行動形態を実際にロボットに指令し動かすことで，シミュレーションによって得られた学習結果を実機に還元することが可能となる．

4･9　まとめ

本章では，強化学習を用いてロボットに前進行動を獲得させる方法について述べた．学習に際しては，ロボットの疲労や学習時間の問題を考慮し，実機から得られた報酬情報を元に 2 次元の配列からなる報酬データベースを作成し，それらを用いて Off-Line で学習を行い，その結果を実機に還元するという手法を用いた．

第5章
学習結果
～人間の報酬操作による機械学習支援～

5･1 概説

本章では，PSD センサから得られた距離情報に基づいた客観的な値を報酬として与えることでロボットが前進行動獲得が可能であることを，シミュレーションおよび実験面から示す．また本章では報酬として PSD センサ出力値そのものを用いた場合と，PSD センサ出力値に各種操作を加えた場合との学習結果の比較検討を行っている．

5･2 報酬情報

4 章で述べたように，学習に当たっては報酬情報をあらかじめセンサにより取得しておき，それらの値を参照しながらシミュレーションを行う．

ここで本研究で用いる報酬情報を示す．

Fig.5-1 に示すようにある状態 S_t から次状態 S_{t+1} へ遷移した際の移動距離を PSD センサから取得した．ロボットが取りうる状態のパターン数は 25 通りであるから，報酬の数は 625(25×25 のマトリクス)となる．

Fig.5-2 に PSD センサから得られた報酬情報をプロットしたものを示す．X 軸がある現在の状態 S_t を表す．そして行動を取ることにより，ある状態から次の状態へ遷移するわけだが，遷移した先の状態 S_{t+1} を Y 軸として取っている．そして Z 軸を報酬としてプロットしてある．つまり例えば 3 という状態から 14 という状態に遷移した際の報酬は（X,Y,Z）=（3,4,Z）= （3,4, r_t [3][4]）というように参照すればよい．

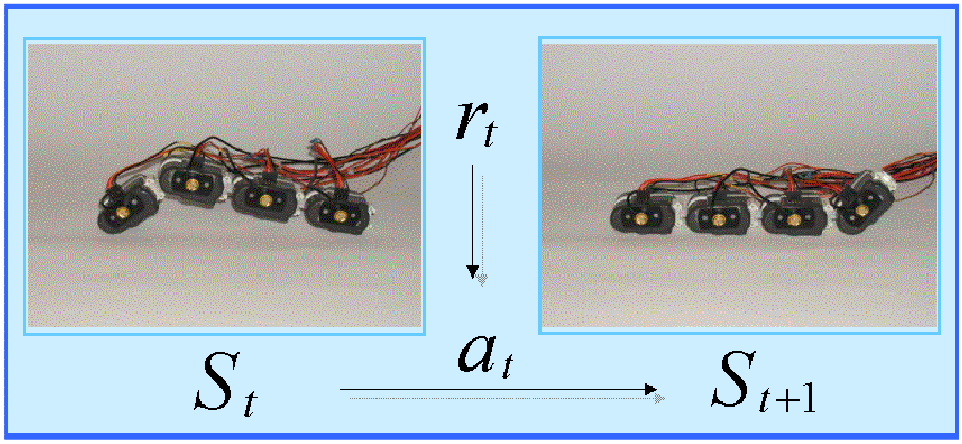


Fig.5-1　Experiment Systems

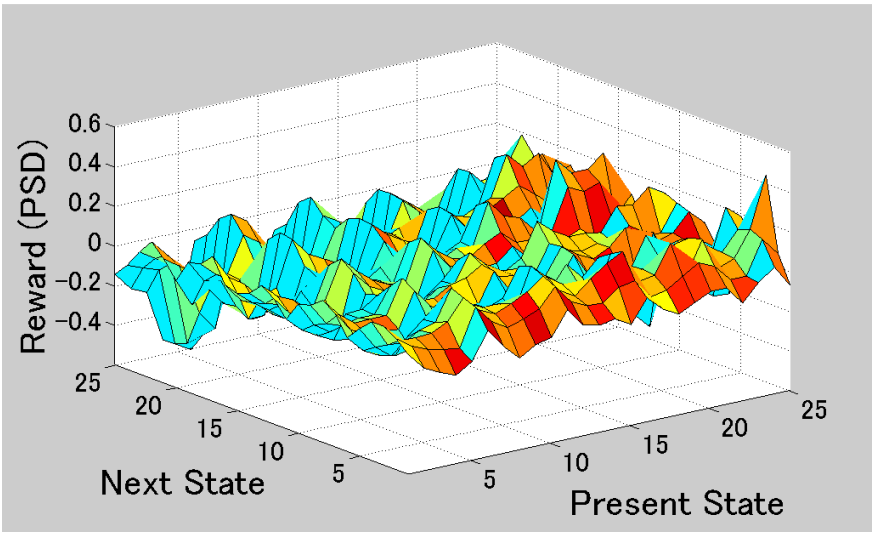


Fig.5-2 Reward information

Table 5-1 Amount of characteristic

Average value	-3.0×10^{-3}
Maximum value	4.4×10^{-1}
Minimum value	-5.7×10^{-1}

次に Table.5-1 に報酬の平均値，最大値，最小値を示す．報酬値は PSD センサのアナログ電圧出力値そのものであるが, 4 章で述べたようにアナログ電圧出力値と移動距離は線形関係にある．そこで学習結果を示す際は直観的に捉えやすいように，これらを距離[mm]に換算して示す．

5-3　距離センサ報酬に基づく学習結果

本節では，取得した報酬情報のうち PSD センサから得られた報酬値 R_p を用いて学習を行ったシミュレーション及び実験結果を示す．

5-3-1　前進移動距離の変遷（シミュレーション）

PSD センサから得られた報酬 R_p を用いて Q-Learning を適用し，シミュレーションを行った結果を Fig.5-3 に示す．第 4 章で示した Q-Learning の式(4-19)のパラメータは学習率 α=0.9 および割引率 γ=0.9 の固定値で学習を行った．図は横軸に学習回数(エージェントが 1 ステップ動く行為)を，縦軸にその学習回数の時の行動価値関数を参照して算出した一歩あたりに進む距離を示している．なお図には学習のランダム性を考慮して 3 試行分のデータを載せてある．

いずれの試行においても，学習回数と共に一歩当たりに進む距離が増加しているのがわかる．

5-3-2　行動価値関数の変遷

Fig.5-4 に Q-Learning の学習過程を示すために，シミュレーションにより得られた行動価値関数を学習回数ごとにプロットしたものを示す．一般にニューラルネットワークなどに代表される教師あり学習や，GA といった探索アルゴリズムと同様，強化学習に関してもその学習過程を外から図り知ることは容易ではない．しかし行動価値関数 Q 値の変遷を観察することで，その概略を把握することができる．Q-Learning では学習初期での行動価値関数は各行動を探索した回数が少ないことから即時的な報酬がそのまま価値となるため，行動価値平面が凸凹となっている．しかし学習が進むとともに将来の報酬を考慮したものが価値となるため，行動価値平面がなだらかなものとなり収束していくのがうかがえる．

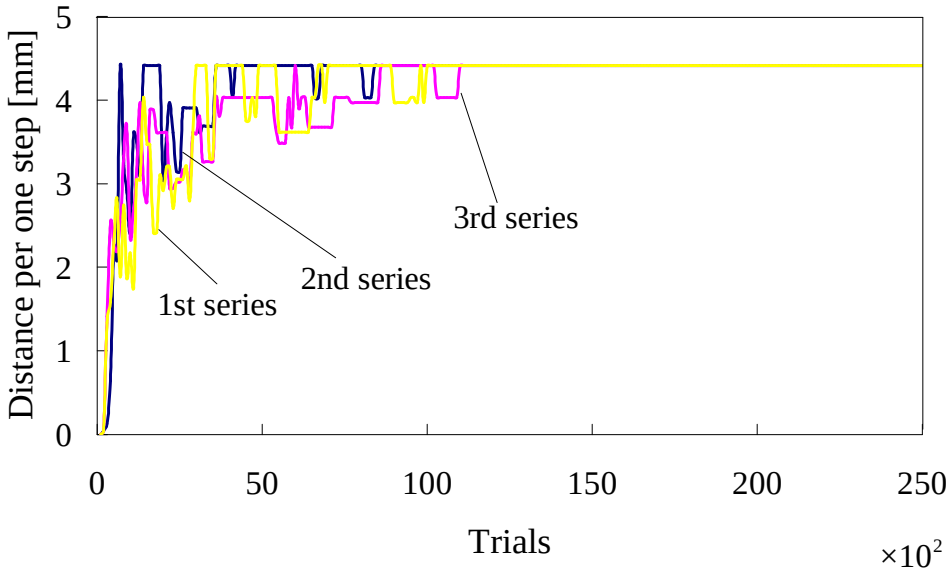
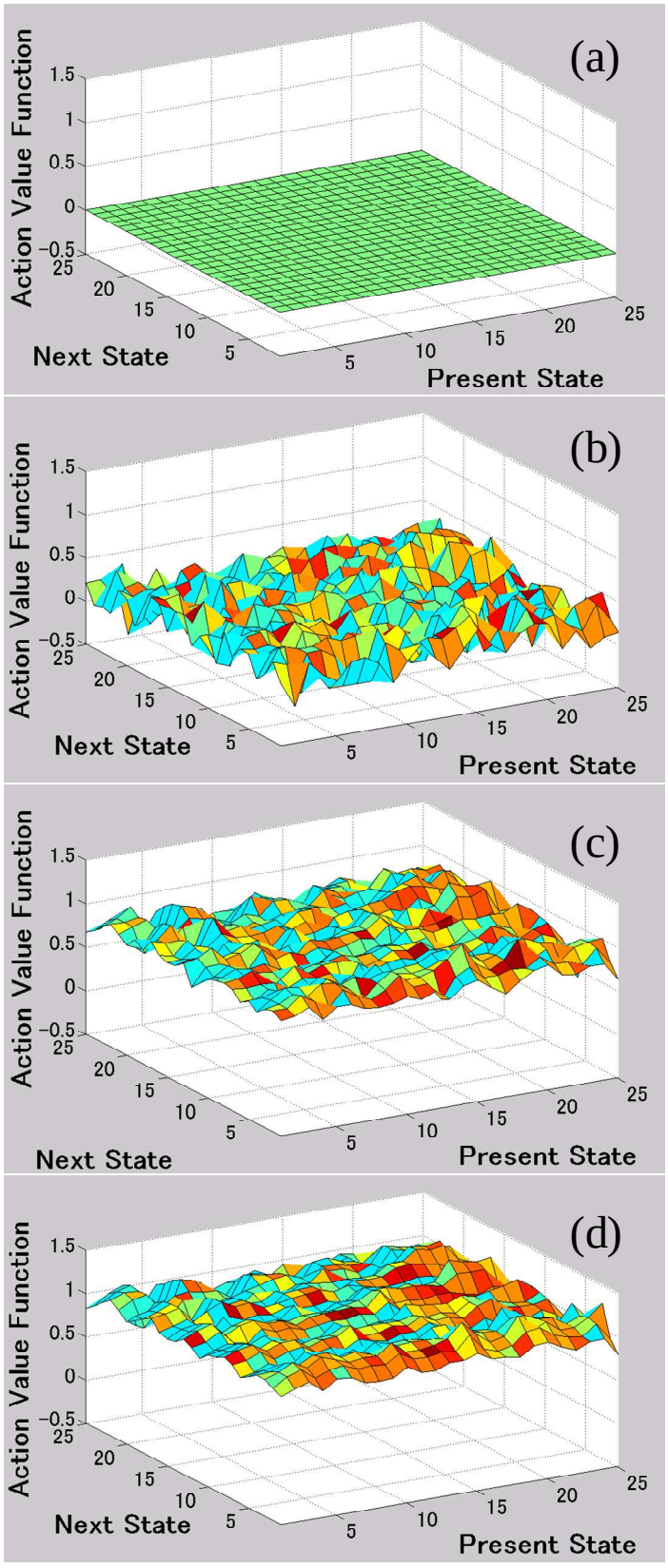


Fig.5-3 Simulation Results



(a) Trials = 0　　(b) Trials = 1,000
(c) Trials = 5,000　　(d) Trials = 15,000

Fig.5-4 Action value function change

5･3･3　前進移動距離の変遷（実験）

シミュレーションでは学習が進むにつれて，より効率よく前進することが出来る行動パターンが形成されていく．それらの得られたパターンを実機に還元し，実験を行った結果を Fig.5-5 に示す．Fig.5-5 からシミュレーション結果と同様に，実験的にも学習回数が増えると共に，より効率よく前進する行動を獲得しているのがわかる．実験では摩擦状態の不確実性から，必ずしも毎回同じ移動距離を得られるわけではないが，ある決まったパターンで動いていることがグラフからも読み取れる．

以上から実ロボットに強化学習を適用し，前進行動獲得が可能であることが示された．

5･3･4　行動形態の変遷

強化学習の本質の部分は，報酬を与えることのみで，学習以前に予想もしなかった行動形態を獲得することである．この行動形態の獲得とは，強化学習のアルゴリズムの特徴として，割引率を用いた将来の行動パターンも考慮して，その中で一連の繋がりからなる最適な行動形態を求めることを可能にしていることである．このように，一連の行動パターンから行動形態を獲得することは，他の学習アルゴリズムに見られない強化学習の重要な特徴である．しかし，この強化学習による行動形態獲得についての報告例は少ない．本項では，学習による前進行動形態の獲得について検討を加え，学習により，どのように行動形態を獲得していくかを明らかにする．本論文での行動パターンは，各関節角の違いから 25 通りの行動パターンがある．行動形態は，行動パターン 12（関節角が全てゼロで，ロボットが伸びきった状態）を初期状態とし，初期状態から始まる Q 値の高い行動パターンを探索していくと，Fig.5-6 に示すように行動パターンの繋がりが出てくる．学習回数が 200 回の時は，最初に予備的な行動があって，最終的に行動パターンの 20⇒18 の繰り返しのループに入ってゆく．（この繰り返しのループを行動形態と定義する．）

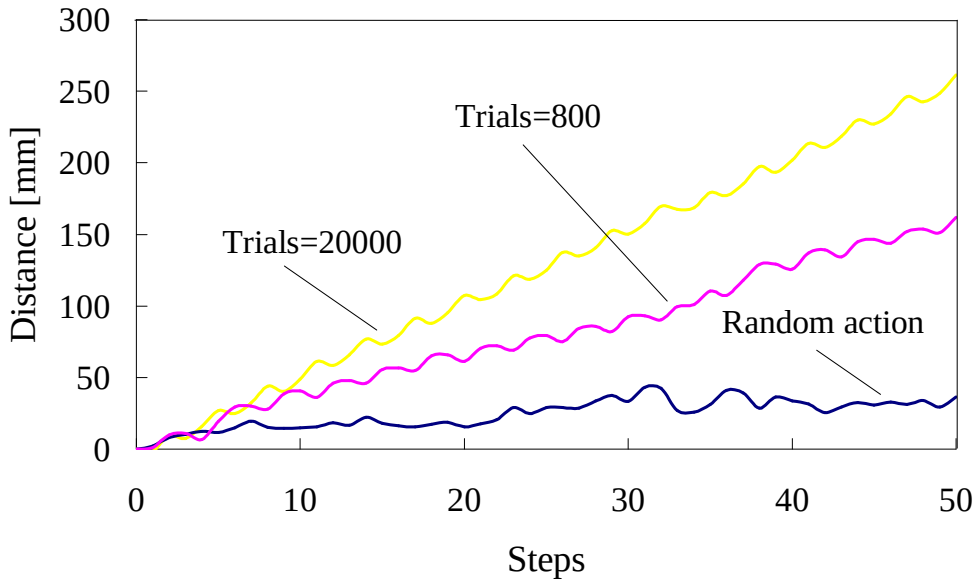


Fig.5-5 Experimental Results

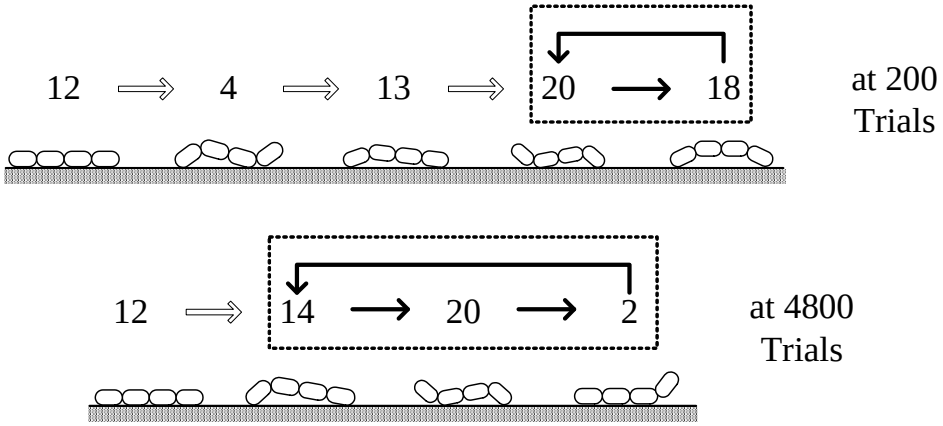


Fig.5-6 Motion pattern change under learning process

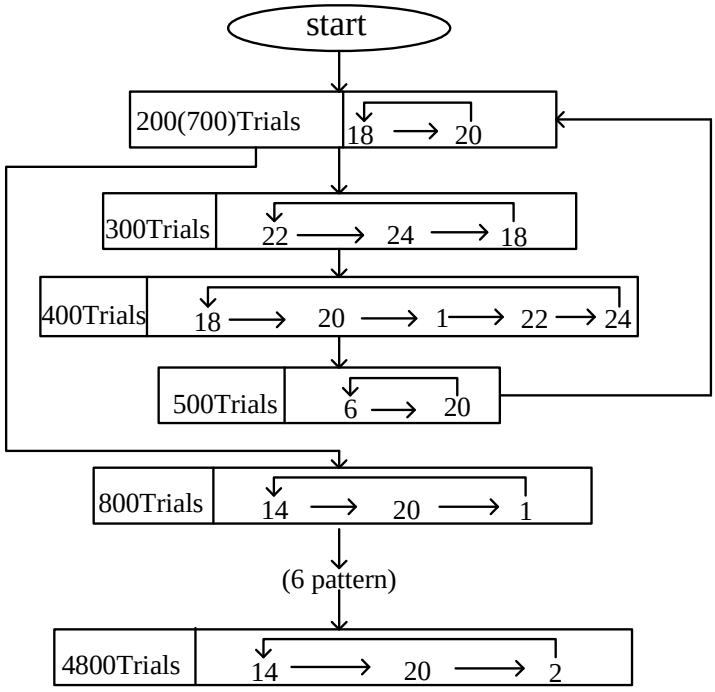


Fig.5-7 Evaluation process of motion form.

学習回数が 4800 回の Q 値の収束状態では，Fig.5-6 に示すように，初期状態からただちに，行動パターン 14⇒20⇒2 の行動形態に落ち着いてゆく．このように，学習が進むと，初期状態から直ぐに最適な行動形態に収束することがわかる．学習プロセスと上記の行動形態の内容について議論を進める．学習率 0.9，割引率 0.8 の行動形態の変遷を Fig.5-7 に示す．この図に示すように，行動形態は紆余曲折しながら最終の最適行動形態を獲得してゆく．行動形態は学習回数 800 回までは，最適な行動形態の探索である．その時の行動形態を模式化したものを参考のために Fig.5-8 に示す．学習回数 200 回では行動パターンの 18⇒20 繰り返しで，この時の行動形態はブリッジ型の形（18）から，それをつぶした形（20）に移す時に前進移動を獲得している．学習回数 300 回の時には，ブリッジ型の形を微妙に変化（22⇒24⇒18）させ，ブリッジの両端の支持ポイントを動かすことで，前進移動を獲得している． その後，学習回数 400 回では，前記の二つの形態を複合した形態に移行し，学習回数 500 回では，頭部の関節を上下させることで，前に這

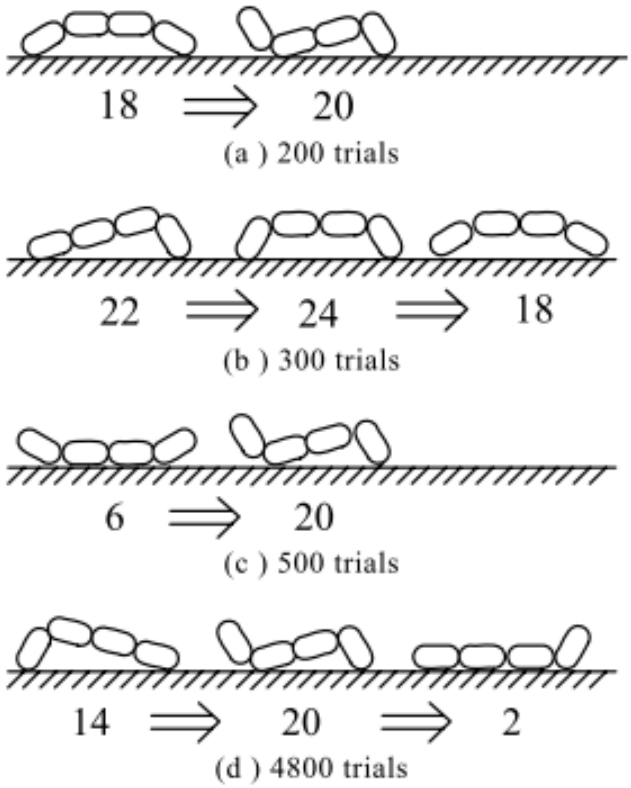


Fig.5-8 Schematic figure of motion form.

うような行動形態で前進行動を獲得している．この後，学習回数 700 回では，学習回数 200 回の同一の形態に戻ってきている．ここまでの学習プロセスは，いろいろな行動形態を発見しながら探索をし，最適な行動形態を探索する学習プロセスとなっており，最終的には，最初に発見したブリッジを押しつぶす形態に帰りつく．次の学習プロセスは，この形態をベースにして，より良い行動形態の探索になっており，学習回数 800 回でより改良した形態を獲得している．すなわち，学習 700 回のブリッジ押しつぶし形態では，押しつぶす過程でしか前進機能を獲得することはできなかったが，改良型ではブリッジの形成時にも前進機能を獲得している．すなわち，Fig.5-8 に示すように，行動形態(a)では，ブリッジの押しつぶし過程で前進機能を獲得するが，ブリッジ形成過程では，前進機能は発現されない．しかし，Fig.5-8 (d)に示すように，新たな行動パターンを挿入することにより，ブリッジ形成過程（2⇒14）とブリッジ押しつぶし過程（14⇒20）の両方で前進行動機能を獲得することができている．この行動形態が最適な形態となる． 学習回数 800 回から Q 値が収束する 4800 回までの学習過程は，上記の行動形態の中で，最適な関節角度を探索するパラメータチューニングの段階になっている．すなわち，学習回数 800 回から 4800 回の間に 6 つの行動形態を獲得するが，多少の揺らぎはあるものの基本的には最適な行動形態の関節角度が若干異なったものとなっている． 行動形態の変遷をまとめると，学習過程の 800 回までは，行動形態の構造獲得探索のプロセスで，学習回数 800 回以後は，最適な構造での最適なパラメータと探索となっている．この学習過程では，人間の学習過程と同様な過程を変遷していると思われ，すなわち，最初の段階では，構造の学習を進め，続いて構造が確定するとパラメータの学習になっている．学習過程に確率的な揺らぎを与えて学習を進めているので，毎回異なった

学習過程を踏まえて最適な行動形態に行き着くが，最終的な最適行動形態はどのような学習過程を踏まえても Fig.5-8 (d)に示す同一のものとなる．また，ここで示した学習過程は，構造学習とパラメータ学習の過程が比較的にきれいに分かれたが，両者が混在する学習過程の場合も見られる．

5･4　報酬操作による学習改善

強化学習においては，「報酬」というスカラー量がロボットに与える情報であり，学習の鍵になっていることがわかる．前節では PSD センサから得られた距離情報に比例した値を報酬として用いて，それぞれ学習を進め，前進行動獲得が可能であることを示したが，本節では報酬に変化を加えてロボットに与えることで，より効率的に学習を進める手法を提案する．

5･4･1　報酬操作の意図

従来強化学習における報酬は，操作されるものではなかった．この背景には恐らく，シンプルな報酬でロボットが学習できることにこそ強化学習の利点があり，報酬を複雑に操作することは機械学習の自律性を失わせるものと考えられたためである．しかし，人間の学習では厳しい教師，優しい教師がいて教師との関係で生徒の学習進度が異なるように，強化学習の報酬自体の与え方は重要であると考える．特に次章で述べる報酬の主観性という議論を扱う上でも，報酬を操作することで学習結果が大きく変わる事実は整理・把握しておく必要がある．

また上記のような定性的な表現だけでなく，定量的にも報酬設計の重要性がある．

ここまで述べてきたように，強化学習では報酬というスカラー量を頼りに，累積的な報酬を最大化するためにはどのような行動を取るべきか，ということを学ぶ学習手法である．例えば Fig.4-4 で示した迷路問題においてはゴールに直接結びつく行動（ゴール手前での一歩など）をした時に＋10 の正の報酬を，それ以外の行動では無駄な行動として－0.1 の負の報酬を与えることで，スタートからゴールへ至る最短の経路を学習していった．なぜエージェントはたったこれだの情報でゴールへ行く最短の道筋を学習できるのだろうか？

その 2 つの大きな要素として下記があげられる．

1. 割引率
2. 報酬

1 の割引率については，割引率によって未来にもらえる報酬というものが割り引かれてしまうので，エージェントがいつまでも無駄な行動をしていると，ゴールに到達した時には，「もう遅い」というようにほとんど報酬がもらえないことになる．その為エージェントは何としてでもより早くゴールに到達することで，報酬和を最大化しようとする．

次に 2 つ目の報酬については，エージェントは直接的にゴールに結びつかない行動に対しては，例で言えば－0.1 の報酬を受け取る．つまり無駄な行動をしていくと，毎回毎回－0.1 の報酬を受け取ることになるので，ようやくゴールに到達して＋10 の報酬を受け取った時には，すでに負の報酬が累積していて，トータルでほとんど報酬が得られなかったということも考えられる．その

為エージェントはやはり何としても早くゴールに辿り着こうと学習を進めて行くのである．

強化学習の特徴でもある割引率については非常に重要視され，多くの研究がなされてきたが，与える報酬の問題について扱った研究は非常に少ない．しかし先の迷路問題の例で言えば，無駄な行動を取った時にエージェントが－6 の負の報酬をシステムから受け取っていたら，どれだけ最短で進んだとしてもゴールに到達した時には，トータルではマイナスの報酬和を得ることになる．そうであればエージェントはスタートから動かずじーっとしていることが，最大の報酬和を得られることと判断し，スタートからゴールへ行く道筋を学習することは出来なくなる．

このように，どのような時にどういった値の報酬を与えるかということは，報酬が単純なスカラー量であるとはいえ難しい問題である．強化学習ではあくまで与える報酬は設計者によるものが大きく，エージェント自身が報酬の上位にある評価基準を学習することは難しいことから，設計者が与える報酬という問題については多角的な研究を進める必要があると考える．

そこで本章では報酬を変化させる手法として，以下の 2 つの方法を用いた．

1. 学習初期に On-Off 型の 2 値報酬を与え，段階的に学習を行う（段階報酬）
2. 報酬を累乗することにより，報酬情報を強調させて学習を行う（強調報酬）

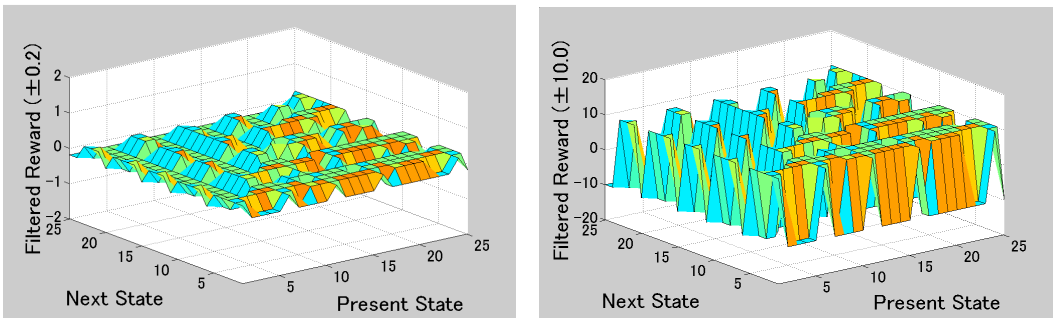
5･4･2　報酬操作Ⅰ:段階報酬

本項では，学習過程において報酬を段階的に変化させることにより，学習の収束性の向上を図った結果を示す．

ここまで報酬として「PSD センサから得た距離に比例した値」を用いていた．これらの距離情報にフィルタをかけてロボットに与える手法を提案する．具体的に言うと，センサから得られた進んだ距離が正の値であった場合，それがどれだけ大きくとも，僅かな量であっても一律に＋○○の報酬を，進んだ距離が負の値であった場合は，やはりそれらの値に関わらず一律で－○○の報酬を，といったように On-Off 型で報酬を与えることを意味する．当然 On-Off 型の報酬を学習初期から最後まで終始与えていたのでは，ロボットは行動の良し悪しを最後まで学習することができない．なぜならば，ロボットにとっては正負の違い以外は全て同じ情報として捉え，どれだけ進むかはわからないためである．そこで学習初期にこのフィルタをかけた報酬を与えることにより，ロボットに行動の大まかな良し悪しを覚えさせ，その後通常の移動距離に比例した報酬を与え学習を継続することで，学習の効率化を図るというのが本節の狙いである．

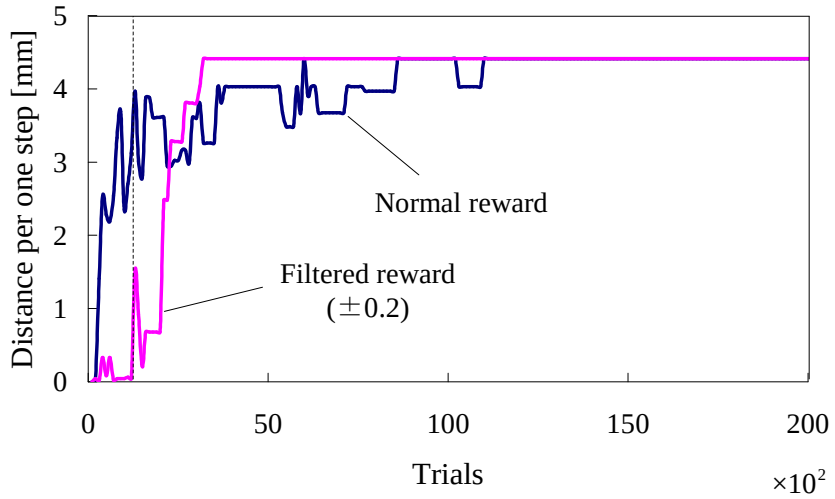
Fig.5-2 および Table.5-1 に PSD センサから得られる報酬の情報を示した．はじめにこれらの報酬にフィルタをかける．ここでは進んだ距離が正であった場合は一律＋0.2 の報酬を，進んだ距離が負であった場合は一律－0.2 の報酬を与える『低フィルタ報酬』と，進んだ距離が正であった場合は一律＋10.0 の報酬を，進んだ距離が負であった場合は一律－10.0 の報酬を与える『高フィルタ報酬』の二つの報酬を用意した．

Fig.5-2 の報酬にそれぞれのフィルタをかけたものを Fig.5-9 に示す．これらの報酬を学習初期に与えて学習を行った結果を Fig.5-10 に示す．図には学習 1000 回までフィルタ報酬を与え，その後は通常の距離に比例した報酬を与えて学習を行った結果を示している．

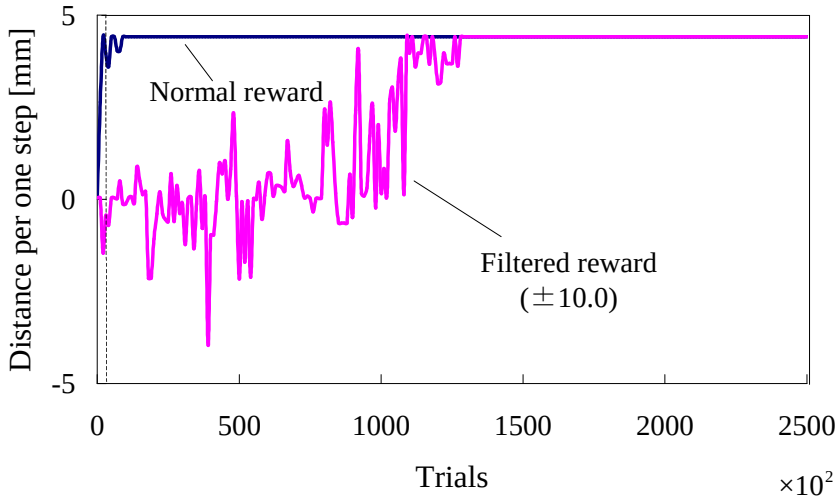


1) ±0.2　　2) ±10.0

Fig.5-9　Filtered reward



(1) ±0.2



(2) ±10.0

Fig.5-10　Filtered reward

図からもわかるように，低フィルタ報酬を学習初期に与えた場合，学習の収束性が大幅に向上した一方，高フィルタ報酬を学習初期に与えた場合，学習の収束性が大幅に悪化してしまったこ

とがわかる．学習初期に与えるフィルタ報酬の値によって改善される場合と悪化させる場合があることから，これらの適切な与え方が存在することが予想される．そこでフィルタ報酬として与える大きさ（絶対値）および，正負の値を違う値で設定した方がいいのかなどについて検討をし，本手法の有効性について議論する．

はじめにフィルタ報酬として与える際，どのような大きさの値（絶対値）を与えるべきかについて議論する．Fig.5-11 に学習初期に与えるフィルタ報酬の絶対量を±0.1～±0.3 で変化させてシミュレーションを行った結果を示す．

シミュレーション結果を見ると，フィルタ報酬として 0.2 で与えると，収束が最も早く最適な値であることがわかる．0.1 の場合では，通常の学習と比べると幾分良好な結果であるが，0.3 に至っては学習の収束が悪化したのがわかる．先に示した Fig.5-10(2)の結果と共に参照すると，0.3 以降その値が上がるほど収束性に悪影響を与えていることがわかる．このように学習回数 1,000 回までは前に進んだら一律の正の報酬，後ろに進んだら一律に負の報酬というように，0.1～0.3 と値は違えど，同じ報酬を与えているにも関わらず，その後（学習回数 1,000 回以降）の学習結果が大きく異なるのは何故だろうか？この原因を解明するために学習途中の行動価値関数の値を見てみることにした．

Fig.5-9 に示したフィルタ報酬を学習回数 1,000 回まで与えて学習を行った際（学習結果 Fig.5-10）の行動価値関数の変化を Fig.5-12 に示す．

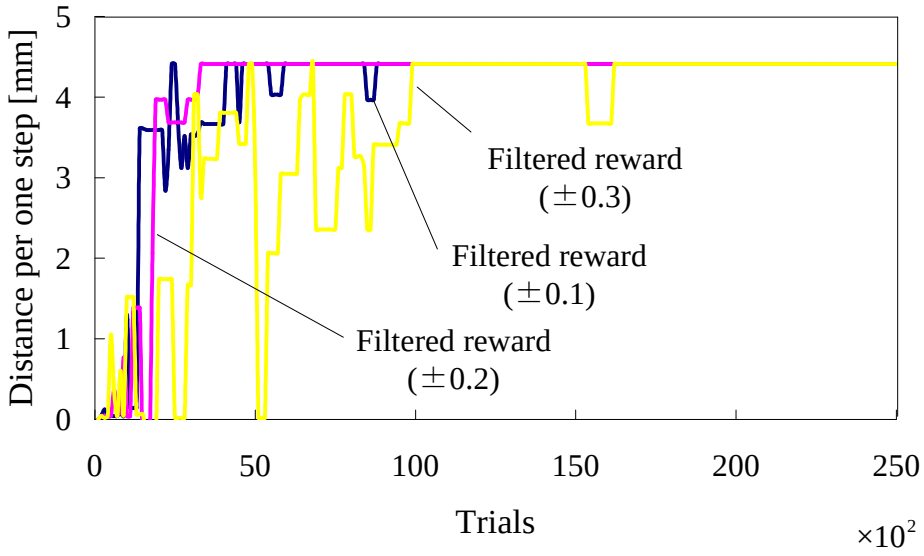
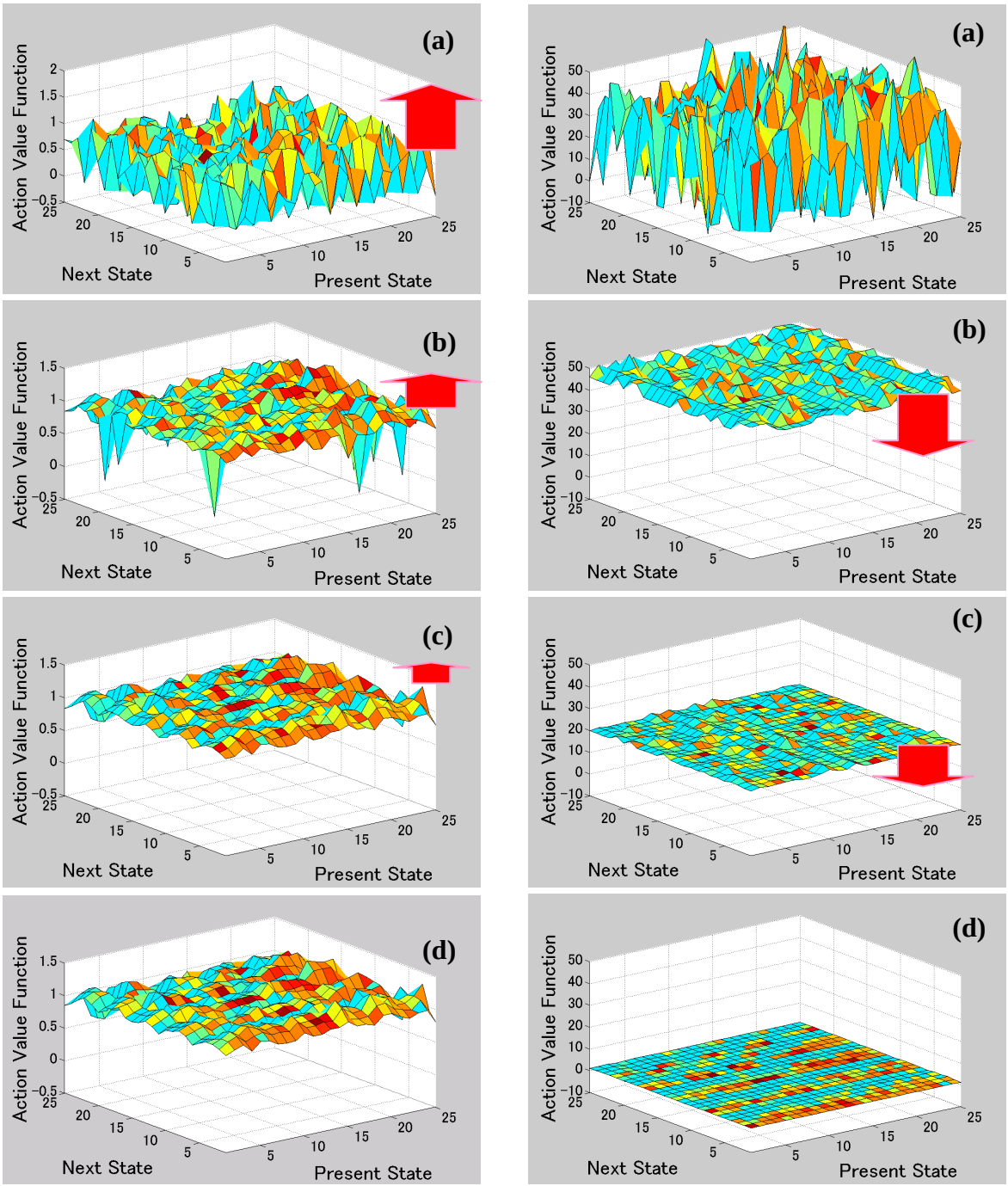


Fig.5-11 Simulation results



(1) Filtered reward =±0.2

(a) Trials=1,000 (b) Trials=3,000

(c) Trials=5,000 (d) Trials=12,000

(1) Filtered reward =±10.0

(a) Trials=1,000 (b) Trials=5,000

(c) Trials=30,000 (d) Trials=150,000

Fig.5-12 Action value function change

行動価値関数の変化を見てみると，±0.2 のフィルタ報酬を学習初期に与えると，与え終わった学習 1,000 回の時点で行動価値関数が 0.0～1.0 のあたりにばらつき，その後の学習で行動価値関数が下から上へ収束していくのに対し，±10.0 のフィルタ報酬を学習初期に与えると，学習 1,000 回の時点で行動価値関数が 0～50 の広範囲にばらつき，学習回数 5,000 回あたりで行動価値関数が 50 付近で一様になる．その後の学習では行動価値関数が 50 付近の値から下へ時間をかけて下がっていくため学習の収束に時間がかかることがわかった．

以上から言える事は最終的に収束する行動価値関数の値を推測し，学習初期の行動価値関数の値がその近傍になるように学習初期にフィルタ報酬を与えることにより，推定値へ短期間で近づけることが，学習の収束性を向上させる効果的な方法であることがわかった．

ここで二つの問題が生じる．

1. 最終的に収束する行動価値関数の推定方法
2. 推定値へ短期間で近づけるためのフィルタ報酬の値

1 の問題に関しては，まず最終収束値を完全に求めるのは不可能であり，もしそれが可能であれば学習は必要ないことになる．その為，おおよその値（どの程度の大きさでばらつくのか）を推測することさえできれば，学習の改善に大きくつながると考えられる．しかし行動価値関数は定義が『累積的に得られる報酬和』であり，割引率などの学習パラメータなどによって，最終的に収束する値が異なり，予測するのは非常に難しい問題である．

次に 2 番目の問題として，仮に大よその最終行動価値関数の値がわかったとして，推定値へ近づけるためには，どのようなフィルタ報酬を学習初期に与えるのかという問題である．

以上二つの問題は解析的に解くのが非常に困難であるが，おおよその値を推定することを本節の目的とする．

はじめに Table.5-1 に示す本システムで得られる PSD センサの特徴量についてもう一度観察して見る．そこで報酬の最大値を見てみると，0.44 である．

ここで行動価値関数 Q という値について今一度振り返って見ると，行動価値関数とはある行動を取った時累積的に得られる報酬和であり，式で表せば以下のようになる．

$$Q_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots\cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \qquad (5\text{-}1)$$

ここで実際には不可能であるが，先に示した最大報酬 0.44(r_{max})というのを常に受けつづけたと仮定する．また割引率 γ は 0.9 とすると，常に最大報酬を受けつづけた場合の行動価値関数 Q_{max} は

$$Q_{\max} = 0.44 + 0.9 \times 0.44 + 0.9^2 \times 0.44 + \cdots\cdots = 4.4 \qquad (5\text{-}2)$$

となる．したがってどのように頑張っても，最終的に収束する行動価値関数の値は 4.4 以上にはならないのである．そしてロボットは複数の行動を組み合わせて，累積的に得られる報酬を最大化しようとするから，当然毎回最大の報酬を得られるわけではなく，無駄な動きなども取り入れつつトータルで最大報酬を得ようとする．結果論で言うと，PSD センサ報酬で学習を行うと最終的に収束する行動価値関数の値は，割引率を 0.9 とした場合，0.6 から 1.4 でばらつくことから，毎回最大報酬を得たときの行動価値関数 Q_{max} の 4.4 とは 1/3 以下の値になってしまうことがわかる．この 1/3 という値はシステムによっても違うだろうが，ここで推定割引係数 d として扱うこ

とにすると，おおよその最終収束行動価値関数の値 $\hat{Q}$ は

$$\hat{Q} = d(r_{\max} + \gamma r_{\max} + \gamma^2 r_{\max} + \cdots\cdots\cdots) = d \cdot r_{\max}(1 + \gamma + \gamma^2 + \cdots\cdots\cdots) = \frac{d \cdot r_{\max}}{1-\gamma} \qquad (5\text{-}3)$$

の付近でおおよそばらつくことがわかる．つまりこの推定行動価値関数 $\hat{Q}$ をあらかじめ一様に初期行動価値として初期化しておき学習を行うことで，最終収束行動価値関数の値にいち早く収束し，学習がスムーズに行われると考えられる．

そこでこの手法の有効性を確かめるために，行動価値関数をあらかじめ上記で定めた $\hat{Q}$ で一様に初期化して学習を行った結果を Fig.5-13 に示す．なおここでは推定割引係数を 0.3 および 0.2 に設定し，推定行動価値 $\hat{Q}$ を 1.32 および 0.88 として学習前に初期化しておくことで学習を行う．

結果を見ると，行動価値を推定せずに全てゼロで初期化した場合(d=0.0)の学習に比べ行動価値関数を推定し学習を行ったほうが，早く収束することがわかる．また推定割引係数 d に関しては 0.2 と低めに取った方がよいことがわかる．結果的には d=0.2 とすることにより，実際に収束する行動価値関数の最小値 0.6 に近い値を推定していたことになる．この推定割引係数の定め方については議論の余地があるが，最終収束行動価値をおおよそ推定し，あらかじめその値で初期化して学習を行うことの有効性が示された．しかし先に示した，学習初期にフィルタをかけて，その後通常の距離に比例した値を報酬として用いる学習(Fig.5-10 および Fig.5-11)に比べると若干収束性が悪いことがわかる．

このことについて考察を行うと，先に Fig.5-13 で示したシミュレーションでは全ての行動価値を一律に推定値 $\hat{Q}$ で初期化していた．つまり学習初期に行動価値の優劣はなく，最終収束行動価値の大きさに近い一律の値で初期化をしていた．しかし，フィルタ報酬を与えた学習では，学習回数 1000 回まで一律の報酬とはいえ，正負の優劣に対してだけは情報を与えているので，学習回数 1000 回の時点で，ある程度ばらつきをもち，なおかつ最終収束行動価値に近い大きさをもつ値

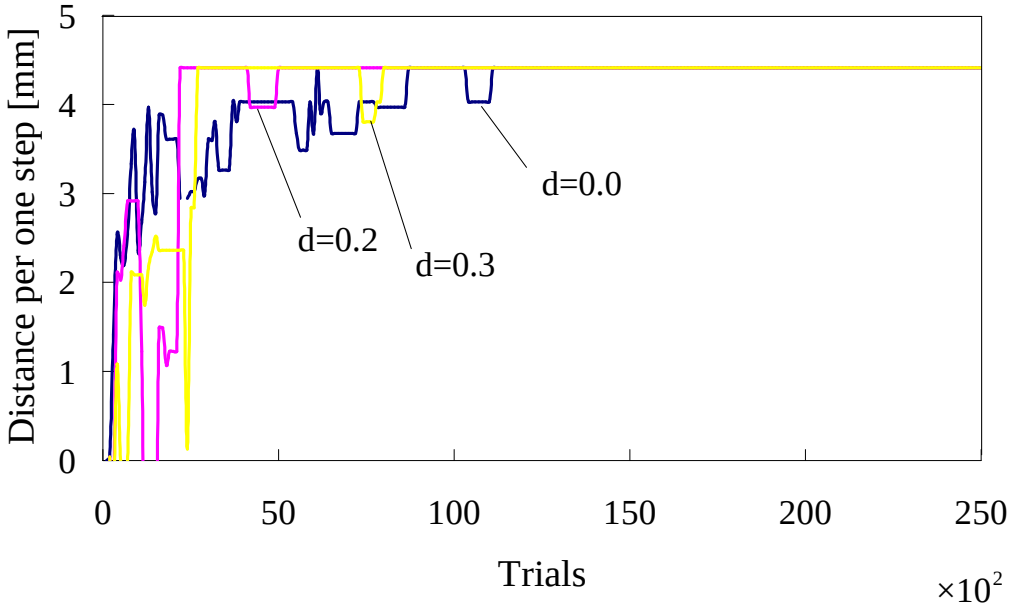


Fig.5-13 Simulation results

が，それぞれに格納されていくことになる．このため，全てを一様に初期化しておく方法よりも，多くの情報を学習初期に得られることになる．

しかしここで問題となるのは先にあげた二つ目の問題である．どのような大きさのフィルタ報酬を学習回数何回まで与えることにより，短期間で推定行動価値に近づくのかという問題である．

はじめに何回まで与えるかという問題についてだが，議論を複雑にしないために学習初期1,000回までと固定して話を進める．これについては議論の余地があるのだが，本研究の場合，ある25パターンの状態から次の25パターンへ移る行動パターンが625(25×25)通りあり，一度はこれらの行動を経験させ学習を行いたいという思いから行動のランダム性を考慮して，1,000回とした．もちろんランダム探索であるから，1,000回行動した時点でも経験していない行動というのもあるのだが，ここでは学習回数1,000回までとした．

次にこの学習回数1,000回までにどのような大きさのフィルタ報酬を与えれば，1000回の時点で推定行動価値関数$\hat{Q}$に一番近づくかという問題であるが，この問題に対しても，先に示したPSDセンサの報酬情報を元に推定することを考える．ここでは二つの指針について述べる．はじめに，Table.5-1で示したように，正の報酬値の平均値r_{ave+}はおよそ0.08という値なのであるが，指標としては，前に進んだ場合毎回およそこの程度の正の値は期待できるという値である．そこでこの正の報酬の平均値r_{ave+}の2倍程度の値（ここでは0.16になる）を，フィルタ報酬として与えることを提案する．これは学習回数1,000回という非常に短い間に，平均値より大きな値の報酬を一律に与えることにより，短期間で推定行動価値関数$\hat{Q}$に近づけようという狙いである．

そして二つ目の指針としては，最大報酬値r_{max}の1/2程度の値（ここでは0.22になる）をフィルタ報酬として与えることを提案する．これに関しても，最大報酬値の1/2という比較的大きな値を学習初期に与えることにより，短期間で推定行動価値に近づける狙いがある．まとめるとフィルタ報酬をr_{fil}とすると

$$r_{fil} = 2 \cdot r_{ave+} \tag{5-4}$$

or

$$r_{fil} = \frac{r_{max}}{2} \tag{5-5}$$

となる．この指標を元にすると，先に示したシミュレーション結果で，±0.2付近のフィルタ報酬を与えた際の学習結果が良好であったことに説明がつく．

次にここまでフィルタ報酬は正負の値で同じ絶対値の値を与えていたが，このことについての考察を行う．Fig.5-14に進んだ距離が正の値であったら一律＋0.2の値を与え，進んだ距離が負であった場合，その値を一律-10，-2，-0.2と変えて学習を行った結果を示す．

学習結果を見ると，与える負の値はあまり学習結果に影響を与えないことがわかる．このことについての考察を行う．式(4-19)で示すようにQ-Leaningでは行動価値関数$Q(s_t, a_t)$は報酬r_tと遷移先の最大行動価値$\max Q(s_t, a)$により定まる．本手法では，学習回数1000回以降は，センサから得られる報酬で学習が行われ，また最大行動価値は，学習回数1000回までに与えられた報酬の正の値で何らかの値が作られるため，学習率が高い場合は，各行動価値は正の高い値を持つものに急速に引きずられる．その為，学習初期に与える負の値はあまり大きな意味をもたないのであ

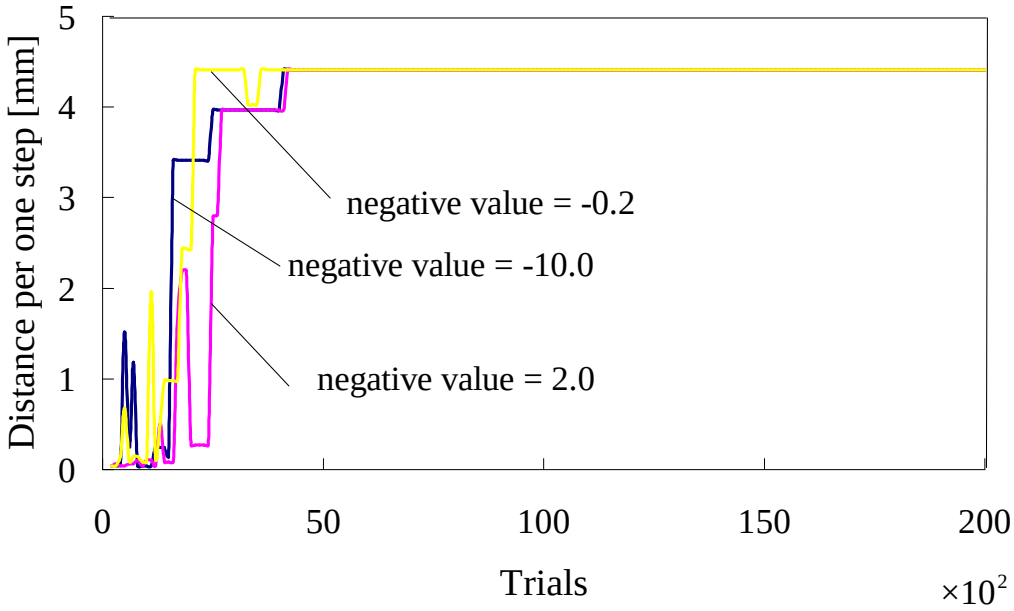


Fig.5-14 The difference of negative value

る．しかし，あまりに小さな負の値の場合，その影響が懸念されることから，正の値と同程度の値で与えれば問題無いと考えられる．

以上の議論をまとめると，収束性を向上させる方法として二つの手法を提案した．

一つは，最終収束行動価値関数の値を推定し，それらの値を一様に学習前に初期化して与えて学習を行う方法．この手法では，推定割引係数 d(0.2～0.3 が望ましいと考える)を定める必要がある．

そしてもう一つは，正の報酬の平均値 r_{ave+} の 2 倍および最大報酬値 r_{max} の 1/2 程度の値を学習初期 1,000 回まで与え，推定行動価値に短期間で近づける方法である．この手法の場合，上記の一様に行動価値を初期化する方法に比べ，短期間で多くの情報を行動価値に集約できる可能性がある．

なお本手法は，Sony 社製 AIBO の前進行動獲得など，複数のロボットシステムについても有効性が確かめられている[220]．

5･4･3　報酬操作Ⅱ：強調報酬

本項では報酬を変化させる 2 つ目の手法として，報酬値を累乗して与える手法を提案する．報酬を累乗するという行為は，例えば Fig.5-15 に示すように，それぞれの報酬値の情報に強弱をつけることになる（負の報酬は絶対値を二乗してから負の値にする）．報酬を強調することにより，ある状態からある状態に遷移する際に得られる報酬の特徴を大きなものとして捉えることで，エージェントが学習しやすいのではないかと考えたためである．

Table.5-1 で示したように，PSD センサから得られる報酬は $-1 \le R_p \le 1$ である．$|R_e| \le 1$ の値では，これらを $R_p^{\,n}(n \ge 1)$ するとそれぞれの値は小さくなるが，値の比率でみれば，大きい値はより

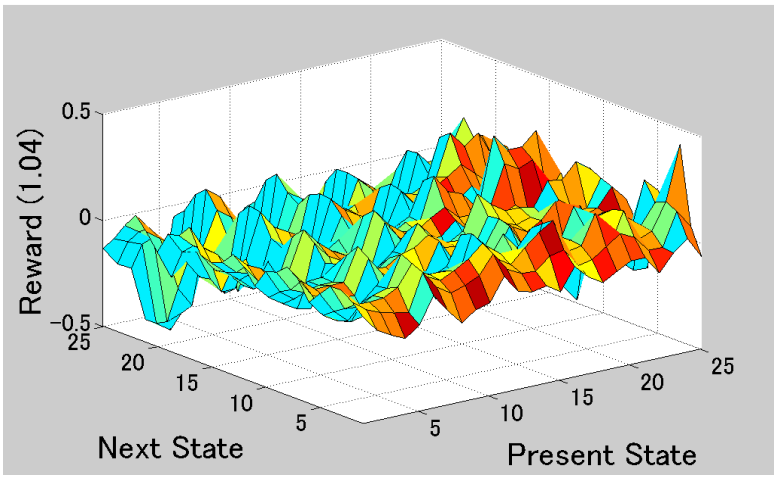
Fig.5-15 Enhancement of reward

大きく，小さい値はより小さくなり，各報酬量の特徴が強調される．そこで報酬を 1.04 乗及び 1.40 乗し，これらの値を学習に用いることにした．Fig.5-16 に報酬を 1.04 乗，1.40 乗したものを視覚的に示す．またそれらを用いて行った学習結果を Fig.5-17 に示す．

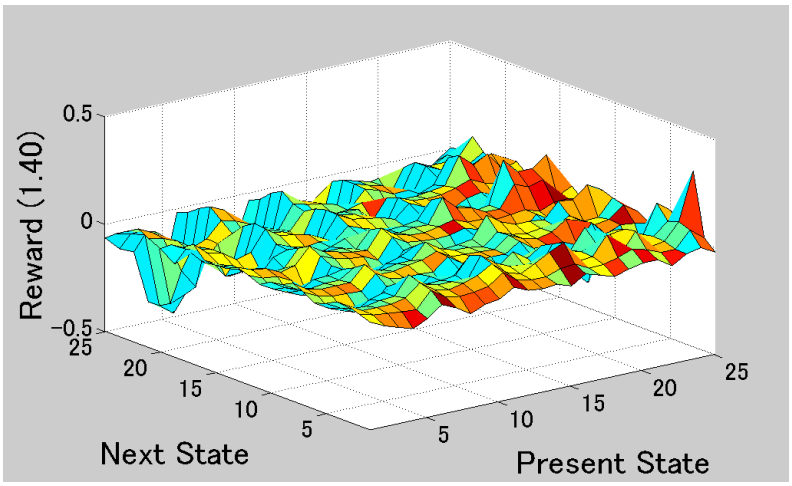
Fig.5-17 から，報酬を 1.04 乗した場合，通常の学習と比較して学習の収束が早くなっている一方，1.40 乗した場合，最終収束値が悪化してしまったのがわかる．

この結果についての考察を行う．報酬を累乗させることは各行動に対する報酬に強弱をつけることになり，それぞれの行動で得られる報酬に大きな差が生じることになる．さらに指数を大きくし累乗することで，報酬の差をさらに広げていくと（$|R_e| \leq 1$ の値では限りなくゼロに近づきながら，お互いの差が大きくなっていくので，プログラム上桁落ちなどに注意が必要だが），全ての状態行動対の中で最も報酬が高い状態行動を数多く選択することがエージェントにとって最良の選択であることから，結果として一連の行動の報酬を最大にする作用よりは，一つのステップのみでの最大報酬を得る作用が強くなる．これは割引率を下げて学習を行うこととほぼ同等の効果となる．

割引率は高いほど将来の報酬を考慮に入れるため，通常 1 に近い固定値で学習を行うことが望まれる一方で，割引率をやみくもに高めると学習の収束に時間がかかる．このため，割引率をある程度下げることは学習を行う上である意味で合理的な手法であり，それと異なる面から報酬値を累乗する手法により割引率を下げるのとほぼ同等の効果を得たということが言える．（厳密な違いについては後に詳細を述べる）



1) The reward to the power of 1.04



2) The reward to the power of 1.40

Fig.5-16 Enhancement of reward

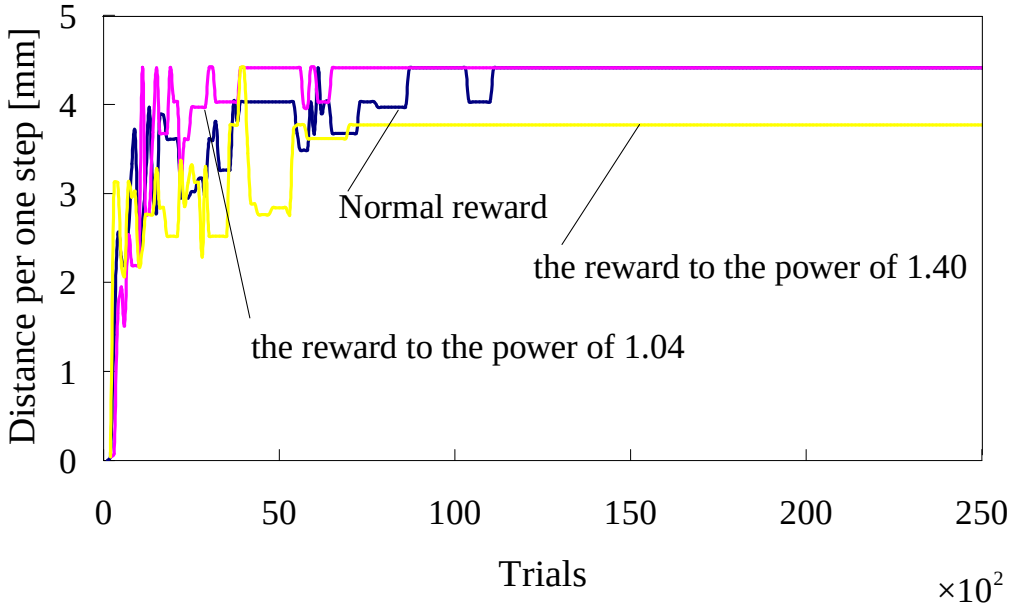
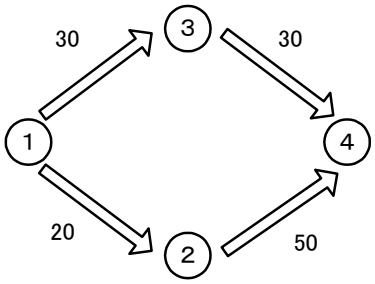


Fig.5-17 Enhancement of reward

しかしこれらの操作は報酬を非線形的に変化させているため，場合によってはそのシステムで得られる報酬情報を壊すことになる．Fig.5-18 にその一例を示す．各行動における報酬が(a-1)のように与えられた場合，トータルに受け取る報酬，言い換えれば価値は①→②→④と進んだ方が高い．一方報酬を二乗に変化(a-2)させても，トータルに受け取る報酬は①→②→④の方が高いため，報酬操作により得られる行動が変化することはなく，報酬値に強弱の差がついた分，割引率を下げたことと同等の効果を得て，学習の収束のみが早まる．一方，報酬が(b-1)のように与えられた場合，価値は①→②→④と進んだ方が高いが，二乗すること(b-2)により①→③→④の価値の方が高くなり，タスクの目標が移動する間に受け取るトータルの報酬を最大化することと考えると，二乗したことにより正しい判断が出来ないことになる．

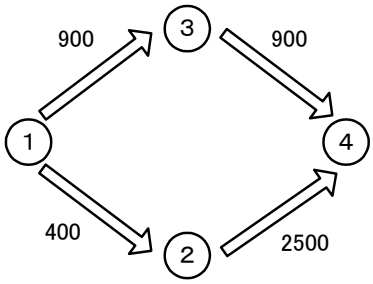
以上の議論を整理すると，Q-Learning は一連の行動の報酬を最大化することだが，本手法の $\max\sum r^n$ の最適化問題の解は $\max\sum r$ の解と必ずしも一致しないことが本手法の問題点となる．今回の場合では，報酬を 1.04 乗した場合では報酬情報を壊すことなく学習の収束を早めることが出来たが，報酬を 1.4 乗した場合では，報酬情報を壊してしまい，最適値に収束しなかったものと考えられる．

以上の結果から，報酬を累乗した値を学習に用いることで学習の収束性を改善させることに成功した．しかし先にも述べたように本手法での問題点は，$\max\sum r^n$ の最適化問題の解と $\max\sum r$ の解が必ずしも一致しない点にある．本研究ではこれらが何乗までなら一致し，学習の改善につな



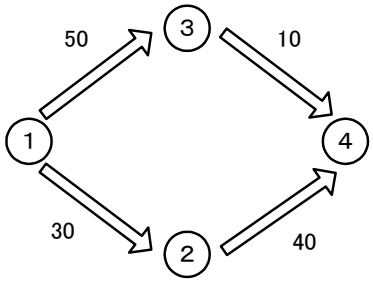
①→③→④　60
①→②→④　70

(a-1) Normal reward



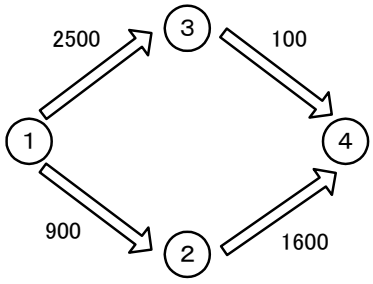
①→③→④　1800
①→②→④　2900

(a-2) Square reward



①→③→④　60
①→②→④　70

(b-1) Normal reward



①→③→④　2600
①→②→④　2500

(b-2) Square reward

Fig.5-18 Reward and value change

がるかなどの明確な指針を打ち出しているわけではない．しかし PSD センサ報酬での学習結果を見る限り 1.04 が最適な指数であり，それ以上の値を用いると $\max \sum r^n$ の最適化問題の解と $\max \sum r$ の解が一致しなくなることから，指数の決定に当たっては極めて小さな値からスタートすることが望ましいと考えられる．現在のところトライアンドエラーで決めているこれらの値だが，報酬値からこれらの値を推定することができると，より一般的に用いることができる手法といえよう．また本手法を実装するにあたって，累乗する際はコンピュータが認識できる桁数に注意して適用する必要がある．

5-5 まとめ

本章では，強化学習アルゴリズムの一つ Q-Learning をロボットシステムに適用し，前進行動獲得が可能であることを示した．ここでは報酬として PSD センサから得られる客観的な値を用いて学習を進めた．強化学習においては「報酬」というスカラー量が学習を進める鍵になっている．従来研究ではこの報酬は○or×や，移動距離に比例した値などシンプルに与えることが一般的であったが，報酬を人間が適切に操作することで，通常の報酬による学習と比較して学習の収束性などを改善できることについて述べた．

はじめに，学習初期に On-Off 型のフィルタ報酬を与えることにより，最終収束行動価値に近い推定値かつ，ある程度のばらつきを持った行動価値を学習初期に配置することができ，より効率的に学習が進むことについて述べた．次に報酬値を累乗したものを与えることにより学習の収束性を向上させる手法について述べた．この手法では割引率を下げて学習を行うことと同等の効果を得られることを示したが，本手法の問題は報酬を非線形的に変化させているために，システムの報酬情報を壊すという点にあった．

第 3 章で，ニューラルネットワーク学習において人間が得意とする「大枠を定める能力」と，機械が得意とする「最適化問題を処理する能力」を組み合わせることで，優れた学習結果が得られることを示した．本章においても「報酬操作」という切り口で，人間が機械の学習に介入することで学習が改善されることを示した．

次章では報酬の主観性について議論を進める．

第6章
機械学習における教示の新たな視点
～教示の主観性･客観性～

6･1　概説

本章では，前章までに述べた複数のロボットシステムに適用した複数の学習手法を比較し，機械学習における人間と機械との関係を整理すると共に，新たな学習の枠組みである教示の主観性・客観性について記述する．

6･2　教示の主観性･客観性

Fig.6-1 に前章までに示した 2 つのロボットシステム及び学習手法を比較した図を示す．

第 3 章では，フレキシブルアームロボットの振動抑制学習を通じて，機械学習における人間と機械の適切な役割分担を述べた．人間が得意とするタスクの大枠を把握するスキルを活用し，機械へ教示を行うことで機械の探索空間を限定し，限定された探索空間で機械が得意とする最適化問題を解くスキルを活かすことが，人間と機械の相互の長所・短所を補完しあう好ましい関係であることを述べた．また教示するデータには人間の主観性（自分ひとりの考え方や感じ方）が含まれていることも述べた．

一方，第 5 章では，イモムシ型ロボットの前進行動獲得学習を通じて，報酬というシンプルなスカラー量を情報として与えることで，機械が行動という上位の機能を学習することを述べた．また報酬を人間が適切に操作，介入することで，機械の学習が促されることを示した．

両者の学習は対象ロボットシステムも，適用学習アルゴリズムも異なる．両学習手法は一般には教師あり学習としての N.N.，教師なし学習としての強化学習という形で，学習アルゴリズム的側面から比較・分類されることが多く，この観点からは全く別物の学習手法とも言える．しかしながら，人間が与えた教示情報を基に機械が最適化問題を解いて望ましい動作を獲得するという点においては，両者の学習に本質的な違いはないことがわかる．そこで機械学習をアルゴリズム的分類ではなく，別の視点で捉えることを考える．それは教示の**主観性 / 客観性**という観点である．人間が機械に何かを教える場合，機械の動作に対して，それを評価する人間には感性や，思想，好みといった，教示者特有の特徴がある．これを強化学習の枠組みの中で表現し，学習結果を観察することで，人間と機械の違いを深彫りすることを目的とする．

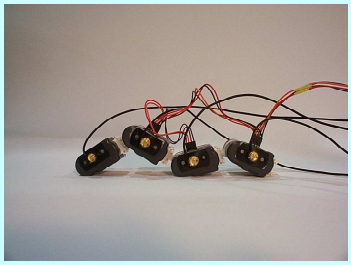
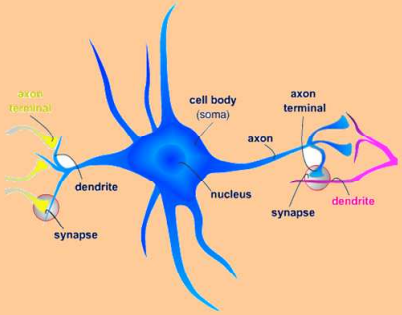
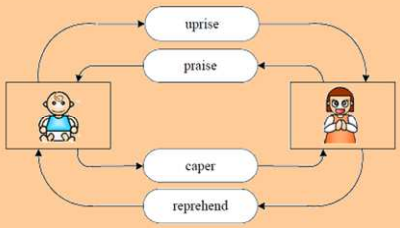
System	Flexible arm robot	Caterpillar robot
		
Learning method	Neural Network	Reinforcement Learning
	Supervised learning	Unsupervised learning
		
Learning objective	Vibration suppression of flexible arm	Acquisition of caterpillar robot locomotion
Teaching	Model structure, Teaching signal	Reward

Fig.6-1 Comparison with each machine learning.

6-3 問題設定

主観とは客観の対義語であり，広辞苑によれば「自分ひとりの考え方や感じ方」である．

Fig.6-2 に強化学習における報酬の主観性・客観性のイメージを示す．第 5 章ではイモムシ型ロボットに与えられる報酬はセンサによる距離という客観的指標であったのに対して，人間により主観的な報酬を与えることを考える．以下距離センサにより与えられる，ものの見方や感じ方に依存しない報酬を「客観報酬」，人間により与えられる，ものの見方や感じ方に依存する報酬を「主観報酬」と定義し，主観報酬と客観報酬とでイモムシ型ロボットの学習結果にどのような差が表れるのか，また評価者が異なる場合，学習結果にどのような影響を与えるかについて考察する．

なお「報酬」と「学習結果」の関係を純粋に比較するのが目的である為，実験での不確実性を排除する為に 4 章で述べた実ロボットと等価の Fig.6-3 に示すロボットシミュレータを構築した．構築方法の詳細は次節以降に述べる．

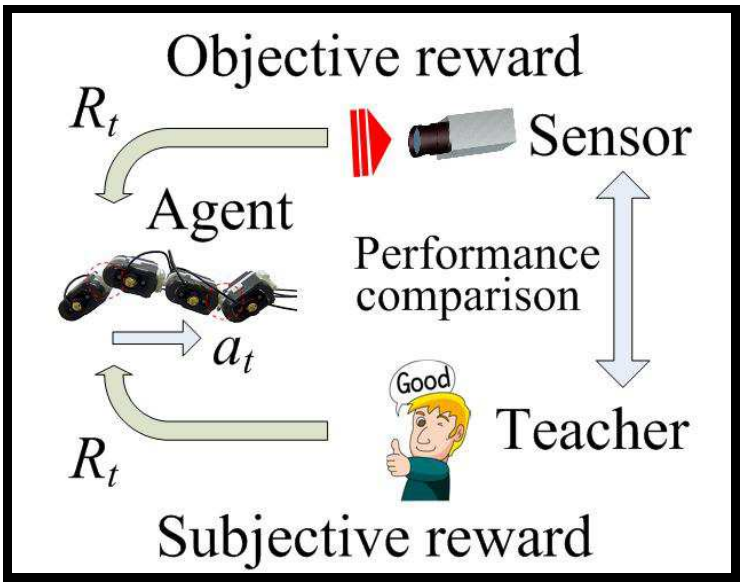


Fig. 6-2 Objective reward and subjective reward

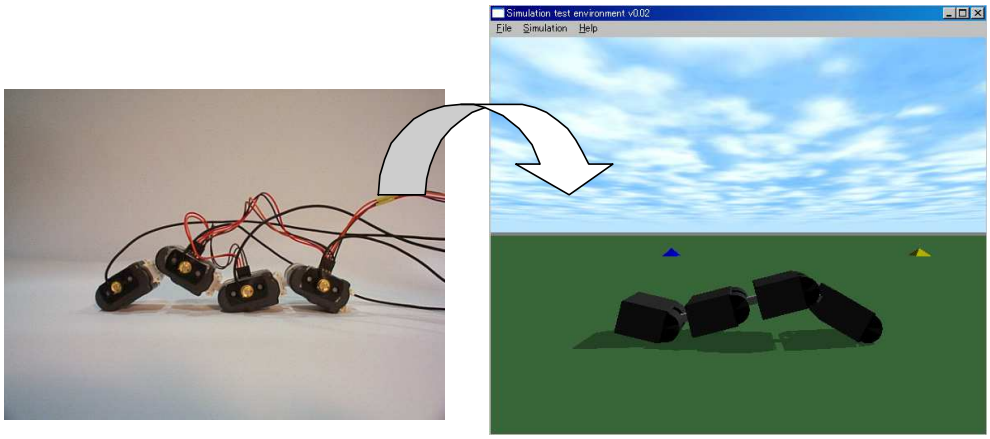


Fig. 6-3 Caterpillar robot simulator

6･4　シミュレーションシステムの構成

6.4.1 Open Dynamics Engine (ODE)

今回用いたシミュレータは Open Dynamics Engine(ODE)を元に作成した．ODE とは，ラッセル・スミス氏らが中心となって開発した 3 次元物理計算エンジンと呼ばれるライブラリである．非常に高いクオリティで物理現象を再現することが可能であることから，広く商用ゲームや研究用シミュレータとして用いられている．高精度な衝突検出まではできないとされているが，床面との摩擦や衝突などの物理現象をライブラリ内の Application Programmable Interface (API)を利用することで実現できる．本項では ODE に関する基礎知識について記述する．

6.4.2 ODE の特徴

主な特徴として以下のことが挙げられる．

i. オープンソース

オープンソースであるため，ソースコードを読むことができ，手を加えることもできる．またオープンソースでありながら，商用レベルの品質を備えている物理計算エンジンである．研究やコンピュータゲーム，3D の製作ソフトやシミュレータの物理計算エンジンとして広く使われている．ライセンスは LGPR と BSD のデュアルライセンスである．詳細は ODE の HP を参照．

ii. マルチプラットフォーム対応

Windows，Linux，Mac でも使用可能．

iii. C/C++言語で開発可能

ODE は C/C++言語で開発されている．そのため，同様に C/C++環境で作成したプログラムを組み込んだり，逆に ODE を C/C++環境で開発されたプログラムに組み込むことが容易である．

iv. 扱いやすい

ODE は初めから 3D のグラフィック機能がついており，簡単な動作テストがしやすい．またマニュアルが充実しており，サンプルプログラムが多い．しかも，コミュニティが活発に活動しているのでメーリングリストにも頻繁にメールが流れ，過去のメールも自由に閲覧できるので情報量が多い．

v. 高速で安定性がある

シミュレーションでは正確性と安定性がトレードオフの関係であり，正確性を重視しすぎると安定性が悪くなり，シミュレーション中で実際の世界では考えられない事態が起きてしまう．このようなことを防ぐために，ODE は正確性より，高速性と安定性を重視している．

6･5 ODE の構成

ODE 中で取り扱われるオブジェクトとシミュレーションの流れについて簡単に記述する．

6.5.1 オブジェクト

i. World

シミュレーションの世界そのもの．World の中にある物体は動力学計算の対象となる．World の中のオブジェクトはすべて同じ時間の流れの中にあり，同じ重量の影響下にある．また，座

標系は直行座標右傾で右手形を取り，単位系は特に決まっていない．ただし，角度だけは[rad]を使用している．

ii. Space

衝突検出に用いる空間．Space の中の物体は衝突検出計算の対象となる．ODE では動力学計算部分と衝突計算部分をわけており，高速化につながっている．

iii. Body

剛体を表す．質量，重心位置，慣性テンソル，位置，速度の情報を持っているが，geometrical な外形の情報は持っていない．したがって，その剛体が球であるのか，立方体であるのかは等にはまったく言及していない. 力やトルクを Body に加えることでシミュレーション内で物体を動かす．そのための関数をメンバとして持つ．

iv. Geom

Geometry の略．衝突検出システムの中で最も基本となるオブジェクトで，Body 単体の形や，Body の集合の形を表す．Body と Geom で剛体のすべての情報を表すことができる．Sphere・Box・Plane・Cylinder・Capsule・Ray や複雑な形状を，ポリゴン・三角パッチで表す為の TriMesh 等が用意されている．Geom は Body と関連付けることができ，Body の位置や速度の情報を用いて，Geom が衝突の判定を行う．

v. Joint

ODE の Joint は 2 つの物体の位置や姿勢を一定に保つ拘束となっている．つまり，ODE では Joint は物理的な実体を持たない．また，接触による拘束も同様に考えられ，Joint として扱われている．Joint は 2 つの Body(または Body と World)との間に作成する．現在の ODE では ball(3 自由度の回転式ジョイント)，hinge(1 自由度の回転式ジョイント)，universal(2 自由度の回転式ジョイント)，slider(直動式ジョイント)，hinge2，contact(接触ジョイント)，fixed(固定ジョイント)等の Joint がサポートされている．また Hinge や Universal にはトルク，Slider には力を加える関数も用意されている．

6.5.2 シミュレーションフロー

ODE による一般的なシミュレーションの流れは以下のようになっている．

Step 0. 描画の準備

視点や視線の設定，表示に使用するテクスチャファイルの場所の指定，シミュレーションループ関数，キー操作関数などの設定をする．

Step 1. 動力学計算 World の生成と設定

動力学計算される物体が入るための World を作り，その重力等を設定する．

Step 2. 衝突検出用 Space の生成と設定

衝突検出される物体が入るための Space を作り，地面などを生成する．

Step 3. 物体の生成

World や Space に入れる物体を生成する．ODE では動力学計算と衝突計算が別々に実装されているので，動力学計算の対象となる Body，衝突検出の対象となる Geom を生成する．その後 Body と Geom を関連付けることでシミュレーションの対象となる物体が完成する．

Step 4. シミュレーションループ

衝突検出や動力学計算を一定時間間隔(step size)に更新し，シミュレーションを 1 ステップ進める．また，ここで物体への力を加えたり，制御したりなどシミュレーションでやりたいことを記述する．さらに，物体の描画も行っているもっとも重要な部分．

Step 5. Space と World の破壊

シミュレーションが終わったら作成した Space と World を破壊する．

6.5.3 衝突検出

ODE には 3 つの衝突検出用の関数が用意されている．

i. dSpaceColide

同じ Space に属する Geom 群の中で接触しそうなペアを探して nearCallback 関数を呼び出す．実際の衝突の検知，接触点の生成等は行われない．

ii. dSpaceColide2

dSpaceCollide と同様のことを異なる Space に含まれる Geom に対して行う．

iii. Colide

実際に 2 つの Geom が接触しているかどうかを調べ, もし接触しているならば dContactGeom オブジェクト(接点)を作成する．DContactGeom には接触点座標，法線ベクトル，めり込み深さ，接触している 2 つの Geom の ID 等が格納される．

dCollide 関数を使うときには, 対象となる Geom のペアを指定して衝突をチェックしなければならないが，すべての Geom のペアに対して衝突をチェックするのは無駄が多い．それを避けるために Geom を Space に追加しておいて，Space に dSpaceCollide，dSpaceCollide2 をかける．それにより，ピックアップされた接触していそうなペアのみに nearCallback 関数の中で dCollide 関数を使って衝突をチェックして衝突の検出を行う．

6-6　生物型ロボットモデルの作成

ODE には Body と Geom のオブジェクトがあり，それぞれ質量特性，座標位置及び外形の情報を保持している．本研究では Fig.6-3 に示したように，シミュレータ上にイモムシ型ロボットを構築した．これは 4 章で示したイモムシ型ロボットの実機をモデルにしており，質量，動作環境などの各パラメータは過去のデータに基づいている．ロボットが取りうる状態パターンは 4 章で述べた実ロボットと同様である．4 つのモータのうち，駆動する部分を Fig.6-4 に丸印で示す 2 ヶ所とし，Q-Learning のテーブル型の離散型データとして扱いやすくするために±52[deg]，±26[deg]，0[deg]の各々5 つの角度を指令値として用いた．このためロボットが取りうる状態パターンは 25 通り(5×5)あることになる．Fig.6-5 にこれら 25 通りの全状態パターンを示す．便宜的にそれらの各パターンに 0～24 の状態番号を割り振った．

以下，報酬の取得方法や学習方法については，第 4 章で述べた実ロボットシステムと同様である為，割愛する．

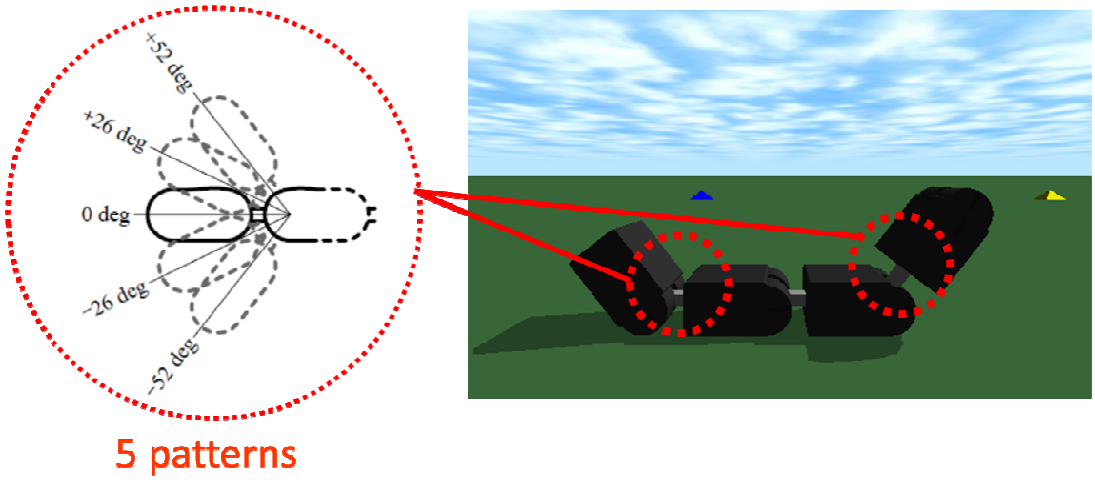


Fig. 6-4 Caterpillar robot simulator

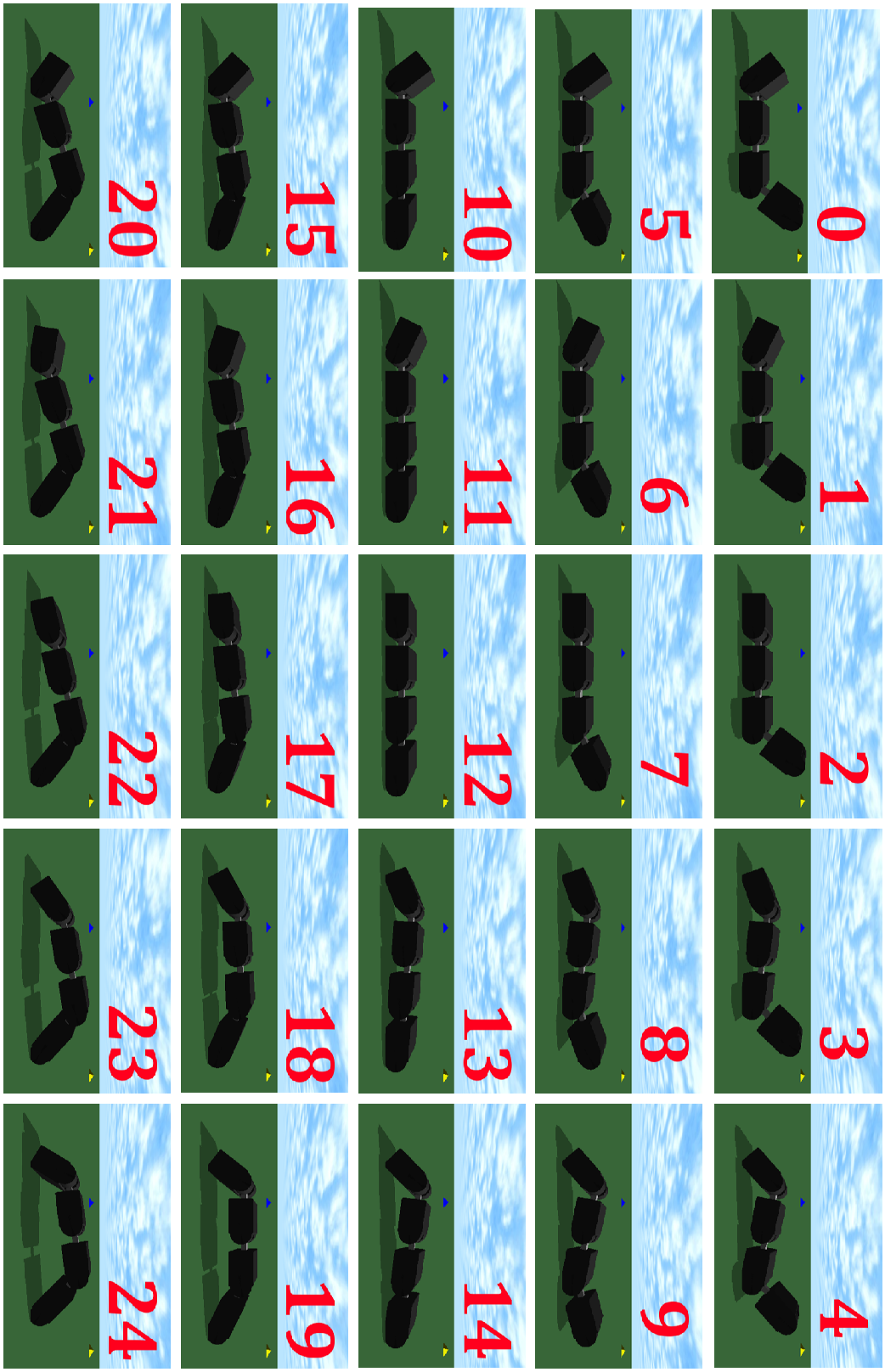


Fig.6-5 Robotic behavior

6-7　評価方法

実験は 5 名の被験者により行った．被験者にはあらかじめ「ロボットを前に進ませたい」という目的のみを教えておく．実験が始まると，ロボットが 625 通りの行動の中からある 1 つの行動を被験者に見せる．これに対して被験者は「非常に良い：+2，良い：+1，どちらでもない：0，悪い：-1，非常に悪い：,-2」という 5 段階で行動に対して評価を与える．行動の良し悪しの評価基準については被験者の主観に委ねるものとする．行動に対する評価を記録すると，ロボットは次の行動を被験者に見せる．これを繰り返すことで 625 通りの行動すべてに対して被験者による評価を与え，主観的な評価による報酬データベースを作成する．なお，見せる行動の順番はすべての被験者で共通である．

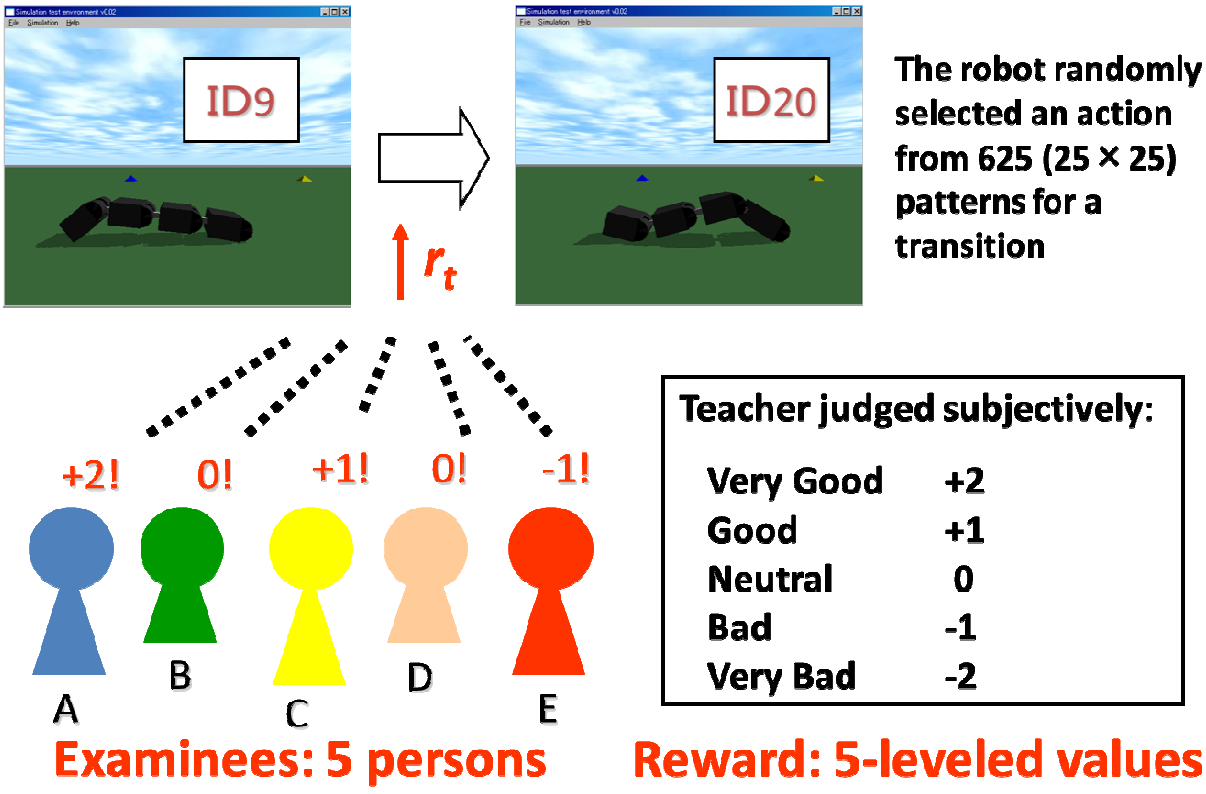


Fig. 6-6 Experimental methodology based on subjective reward

6-8　まとめ

本章では，前章までに述べた複数のロボットシステムに適用した複数の学習手法を比較し，機械学習における人間と機械との関係を整理した．また機械学習における教示の新たな視点である教示の主観性・客観性について，そのコンセプトと検証方法を述べた．

第7章 学習結果Ⅲ
～主観報酬を通した人間の教示特性の理解～

7-1 概説

本章では，主観報酬に基づく強化学習の結果を述べる．

7-2 主観報酬情報

前章で述べた実験により，5名の教師による主観的な報酬のデータベースが得られた．

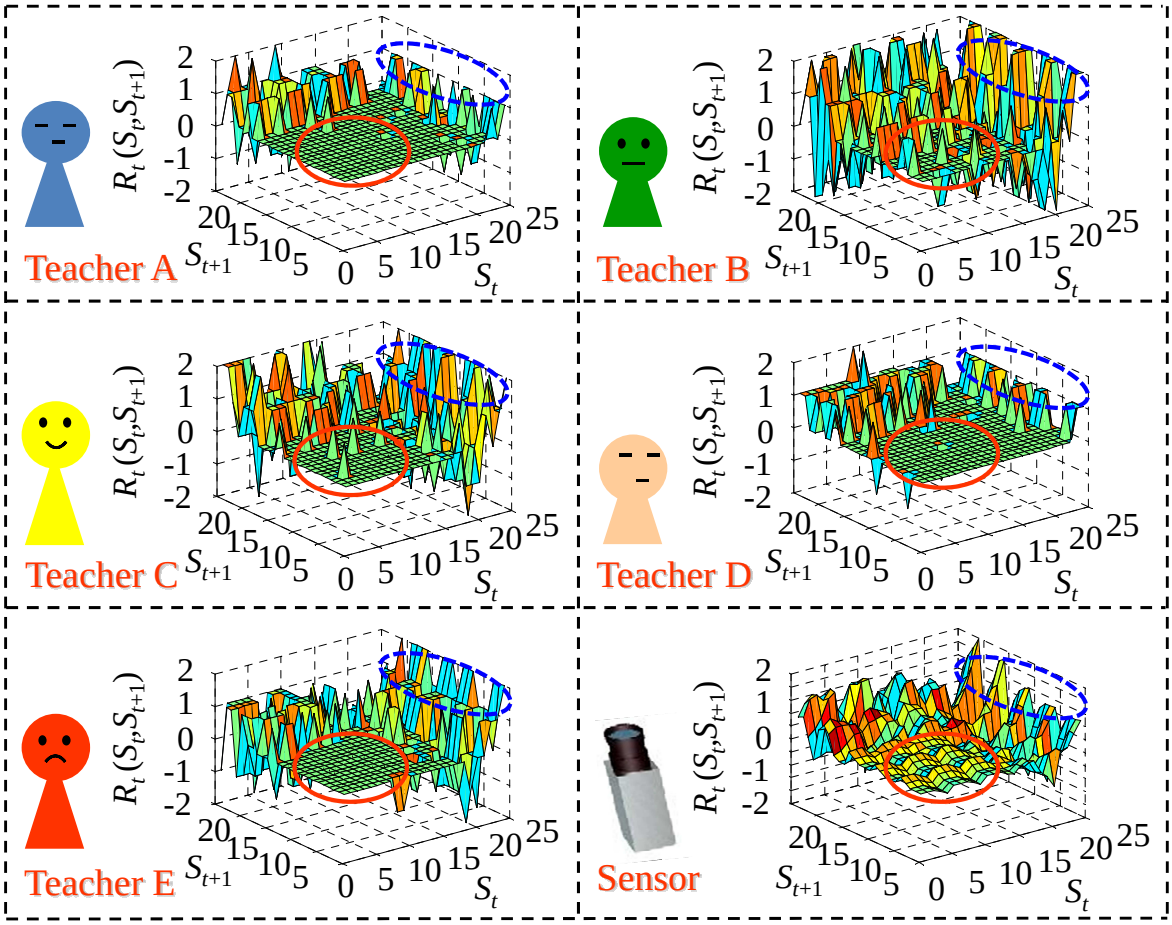


Fig. 7-1 Reward information by each teacher

Table 7-1 Feature quantity for each reward

	Teacher					
	A	B	C	D	E	Sensor
average	0.07	0.12	0.21	0.18	0.06	-0.04
σ	0.46	0.98	0.77	0.45	0.74	0.52

得られた報酬データベースを Fig.7-1 に，特徴量（平均値，標準偏差）を Table. 7-1 に示す．ただし，客観報酬は主観報酬との比較のために，-2 から 2 の間で正規化した．

図表に示すように，同じ状態遷移パターンを見たにも関わらず，各教師が与えた報酬は様々であったことがわかる．これは前進行動を獲得させるという目的に対して，教師が何を教えるべきかという判断基準が，教師ごとに異なることを示唆している．これはセンサによる客観報酬では議論されない報酬の個人差，すなわち主観的教示であり大変興味深い結果である．

7-3 学習結果

Fig.7-2 は各教師の主観報酬及びセンサの客観報酬を用いて獲得された行動形態と，この行動形態で状態遷移を 10000 回繰り返した場合の 1 ステップあたりの平均移動距離を示している．なお複数の行動形態が獲得された場合は，各行動形態の移動量の平均をとった．本システムにおける最適な行動形態は ID2⇒24⇒2 であることが予め分かっている．客観報酬に基づき学習を進めると最適行動形態が得られているが，主観報酬による学習では最適行動形態が得られている教師(A，E)と得られていない教師(B，C，D)にわかれた．また最適行動形態が得られた教師(A，E)についても，単一の行動形態の絞込みはできず，複数の行動形態の獲得となっており，この結果からは客観報酬による学習の優位性が見て取れる．

Teacher	Obtained motion forms	Distance per step [mm]
Sensor	2 → 24	12.9
A	2 → 24 ; 8 → 24	12.7
B	1 → 23 ; 8 → 24	11.8
C	2 → 23 ; 8 → 24 ; 0 → 23 → 2 → 24 ; 0 → 23 → 13 → 22 → 1 → 24 ; ・・・ over 20 patterns	9.4
D	8 → 24	12.6
E	1 → 24 ; 2 → 24 ; 7 → 24 ; 8 → 24	11.9

Optimal motion form **Same motion form**

Fig. 7-2 Learning result

7-4 主観報酬と客観報酬の学習結果比較

主観報酬と客観報酬での学習結果の差について考察する．本タスクの目的は前進移動距離を最大化する行動形態を獲得することにあり，その評価指標は「距離」であり，ものの見方や個人差に依存しない客観的な指標である．この為，客観的な距離情報を報酬として正確に与えられる客観報酬（センサ報酬）での学習結果に優位性があるのは当然の結果とも言える．それでは本タスクにおいては人間の与える主観的な報酬には，距離センサが与える客観的な報酬に対する優位性は無いのだろうか？

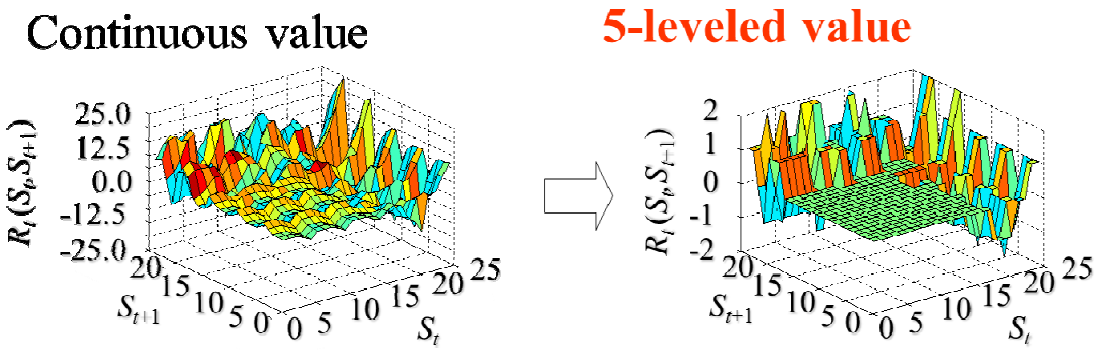
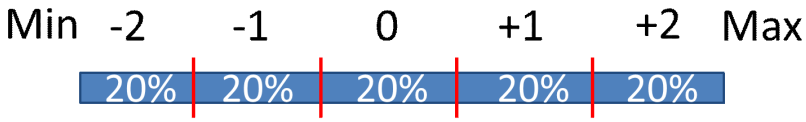


Fig. 7-3 Changing rewards

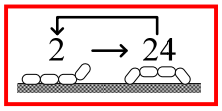
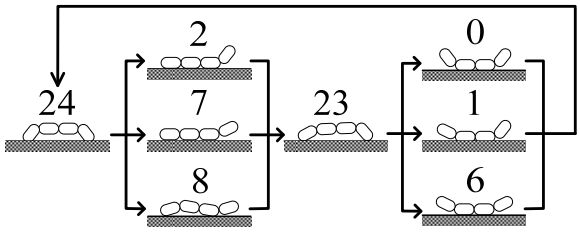


Continuous reward	5-leveled reward
15.50 ≦ R_t ＜ 24.90	2
6.10 ≦ R_t ＜ 15.50	1
- 3.31 ≦ R_t ＜ 6.10	0
12.71 ≦ R_t ＜ -3.31	-1
-22.11≦ R_t ＜-12.71	-2

Fig. 7-4 Range of changing rewards

両者の報酬を情報量と言う観点から比較すると，主観報酬は（-2, -1, 0, +1, +2）の 5 段階の離散値で与えられたのに対し，客観報酬は連続値で与えられていた．報酬の分解能が大きいほど行動の優劣を緻密に教示することが出来るため，客観報酬の方が学習が有利に進むのは当然である．そこで，この報酬の分解能の差を解消するために，Fig.7-3 及び Fig.7-4 に示すようにセンサの報酬をその距離の大きさに応じて連続値から 5 段階の離散値に均等分割し，この 5 段階報酬情報に基づき行動形態を獲得させた．Fig.7-5 に学習結果を示すが，客観報酬を連続値から 5 段階の離散値に分解能を落としたことにより，移動距離が 12.9[mm]から 10.7[mm]へと大きく低下したことがわかる．この要因は分解能が低下したことで報酬による各行動の優劣差の情報が低下した為，ロボットが行動を絞りきれなかった為であり当然の結果と考えられる．しかし最も興味深い点は，Fig.7-6 に示すように報酬の分解能が同じ 5 段階の場合には主観報酬のほとんどが客観報酬よりも優れた行動形態を与える結果となっている点である．これは人間の与える主観的な報酬には単なる距離情報＋αの情報を付加している可能性を示唆しており，非常に興味深い結果となっている．

次節以降では，この主観報酬の客観報酬に対する優位性を多角的に分析する．

Teacher	Obtained motion forms	Distance per step [mm]
Sensor (Continuos)		12.9
Sensor (5-leveled)		10.7

Optimal motion form

Fig. 7-5 Schematic diagram of the learning results – each sensor reward -

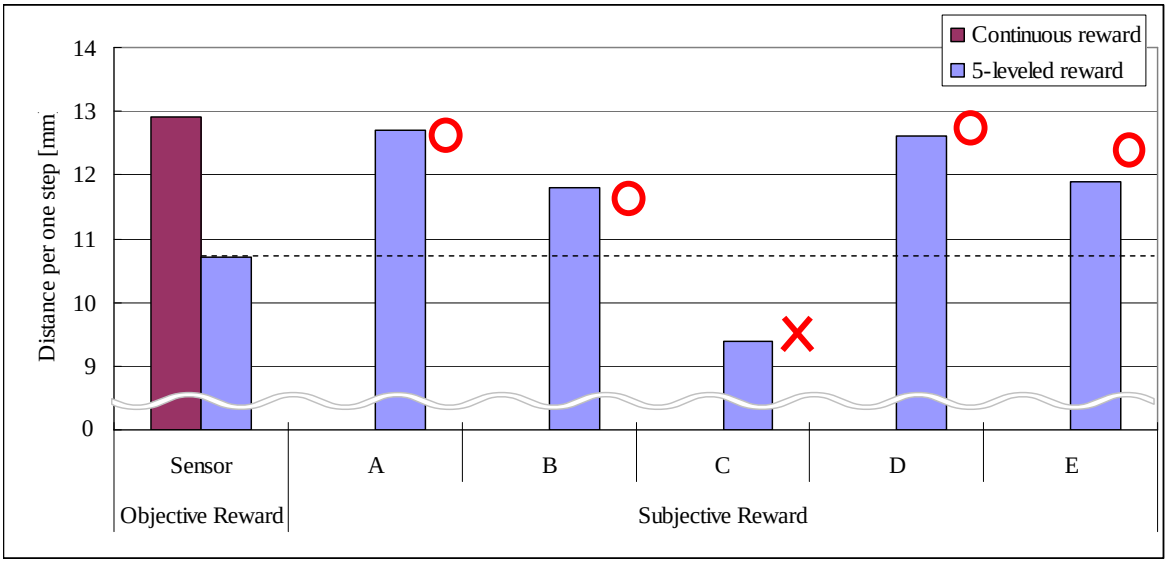


Fig. 7-6 Comparison of the learning results

7･5　問題設定

客観報酬はセンサによる移動距離を報酬として与えているが，分解能への依存度が高い．これに対して人間が与えた主観報酬は，単純な移動量以外にも評価指標がある可能性があり，それによって分解能が制限された状態でも優れた行動形態を獲得させた可能性がある．

この 5 値化された状態における客観報酬と主観報酬の学習結果の違いを「機械と人間の教示の違い」ととらえ，この原因について調査を進める．すなわち，「なぜ 5 値化された報酬において主観報酬が客観報酬よりも優れた行動形態を与えたのか」という問題設定とし，人間の教え方の特徴，及び機械に対する人間の教示の優位性を調べていく．

7･6　人間の優れた教示能力Ⅰ:相対教示

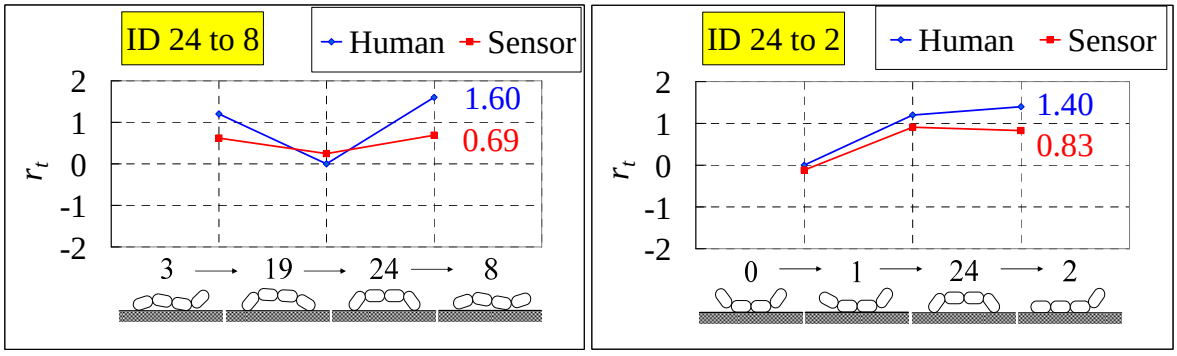
7･6･1　仮説

報酬の分解能が同一の場合，主観報酬の方が良好な学習結果を与えたことに対し，一つ目の原因として相対教示という点に着目した．これは，センサはロボットの現時刻の移動量に応じて報酬を与えるのに対し，人間は過去に観察された動きと比較して相対的に現在の報酬を与えているのではないか，また相対教示には各人に依存した違いが存在するのではないか？と推察した．この推察に対し 24⇒2（最適行動形態の一部）と 24⇒8（全教師から得られた行動形態の一部）の，2 つの主要な状態遷移に注目した．Fig.7-7 は 24⇒2 及び 24⇒8 に各教師が与た報酬の平均値と，その直前 2 回分の行動を示している．ここで，実際の移動量は 24⇒8 よりも 24⇒2 が大きいが，各教師は 24⇒2 よりも 24⇒8 に対してより大きな報酬を与えている．この要因として，24⇒8 の直前に移動量の小さな行動（19⇒24）を見ており，相対的に 24⇒8 の移動距離が大きく見えたのではないか？つまり人間の評価は過去に見た動きに影響を受ける相対的な評価なのではないか？という仮説を立てた．

7･6･2　検証

相対教示の一般性を検証するために，直前の状態遷移に与えた報酬と現在の状態遷移に与えた報酬との間に相関が無いかを調べた．具体的には客観報酬よりも特に優位な結果となった教師 A,D 及び，客観報酬より劣位な結果となった教師 C を特徴的サンプルとして，教師 A,D,C が+1 を与えた状態遷移，及び＋2 を与えた状態遷移の直前の状態遷移に与えた報酬の平均値を，センセのそれと比較した．Fig.7-8 に結果を示す．

Fig.7-8 を見ると＋1,+2 の報酬を与えた状態遷移の直前の状態遷移には，人間はセンサよりも高い報酬を与えていることがわかる．これは「人間は直前に見た移動量が小さいほど，次の動きが相対的に大きく見えて，大きな報酬を与える」という相対教示の仮説とは逆行する結果となっている．また客観報酬に対して優位な結果となっていた A,D の結果と，客観報酬に対して劣位な結果となっていた C の結果との結果を比較しても相関が見られない．つまり相対教示により，主観



(a) State transition ID 24 to 8　　(b) State transition ID24 to 2

Fig. 7-7 Comparison of the reward information

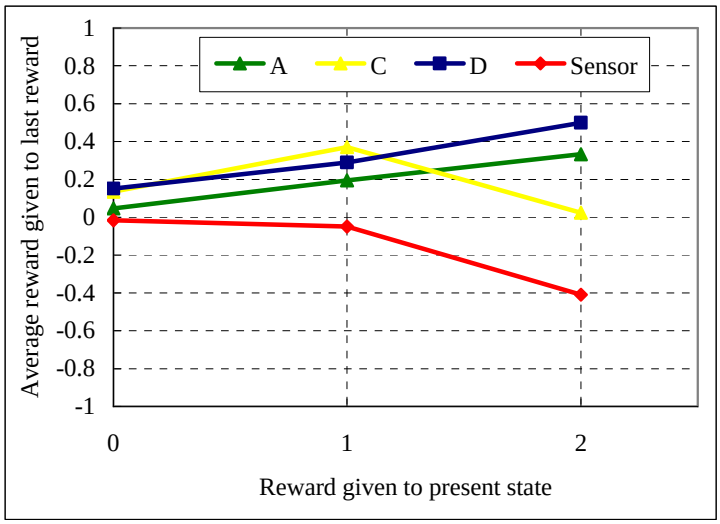


Fig. 7-8 Comparison of the reward information

報酬の方が客観報酬よりも優位な学習結果を生み出した，という仮説は正しくなかったことがわかった．しかしながら，逆に Fig.7-8 からは，人間はセンサよりも，「大きな移動量の行動を見た後には続けて大きな報酬を与えやすい」という傾向が見て取れる．つまり人間は何らか過去の相対的な影響を受けて教示を行っていると言う事実はある．ただしそれがよりよい前進行動獲得につながる有効な教示というわけではないことが明らかになった．

7･7　人間の優れた教示能力Ⅱ：教示分布

7･7･1　仮説

学習を行う上で，主観報酬が客観報酬より優れていた原因の 2 つ目の可能性として，人間が与えた 5 値報酬の分布特性に着目した．これは，報酬は 5 段階と有限であるものの，その与え方の分布特性は無数にあり，この分布が機械と人間では異なるのではないか？また報酬分布は各個人ごとに傾向が異なるのではないか？というものである．ここでも，客観報酬よりも特に優れた行動形態を獲得させた教師 A，D，及び客観報酬よりも結果が劣位であった教師 C（Fig.7-6 参照）の報酬分布に注目した．この 3 人の教師と客観報酬（5 値変換後）の与えた「+2」~「-2」の報酬の個数が，全体の何%であったのかを Table 7-2 及び Fig.7-9 に示す．

注目すべきは，客観報酬よりも結果が優位であった教師 A，教師 D はセンサよりも「＋2」「-2」を与えた比率が低いのに対して，客観報酬よりも結果が劣位であった教師 C は「＋2」「-2」を与えた比率がセンサよりも高かった．同様に「0」を与えた比率は，教師 A と教師 D がセンサよりも高いのに対して，教師 C はセンサよりも「0」を与えた比率が低い．このことから，センサよりも優れた人間の評価の特徴として“「0」を多く与え「+2」，「-2」を厳選して与えている”という可能性が考えられる．これは“普段はあまり口を出さず，非常に良い/悪いものがあったときだけ

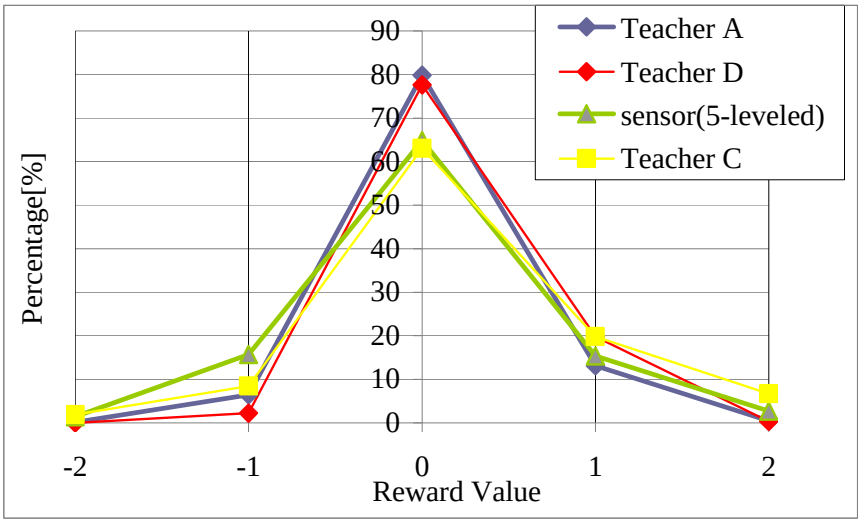


Fig. 7-9 Distribution of rewards

Table 7-2 Ratio of rewards

Teacher	-2	-1	0	＋1	＋2
A	0.16	6.4	79.8	13.1	0.48
D	0	2.24	77.6	19.8	0.32
Sensor (5-leveled)	1.44	15.68	64.8	15.4	2.72
C	1.92	8.48	63.04	19.84	6.72

口を出す”という教え方に相当するものである．また第 5 章 5・4・3 で述べた強調報酬にも類似している．先に示した 5 値化された客観報酬にはこのような分布に特徴は持たせず均等に分割したため，この報酬分布を変化させることで学習結果が改善されるかを検討することにより，「教示分布特性」が人間の持つ優れた教示能力なのかを検証する．

7･7･2　検証

7・4 で客観報酬の分解能を 5 段階の離散値に変換した際，客観報酬の最大値から最小値までの値域を 20%ずつ均等に分割し，それぞれの値域に含まれる報酬に「+2」~「-2」までの 5 値を割り付けていた（Equaled 案）．本章ではこの値域の分割の比率を変更する．Fig.7-10 に分割法の一覧を示す．新たな分割法として，まず「0」の値域を 40%，「+1」，「-1」の値域を 20%ずつ，「+2」，「-2」の値域を 10%ずつとったものを考える（Strict 案）．これは“「0」を多く与え「+2」，「-2」を厳選して与えている”教師 A，教師 D が優位な学習結果を得たという事実を模した分割法である．さらに，この特徴をより強く反映した分割法として「0」の値域を 60%，「+1」，「-1」の値域

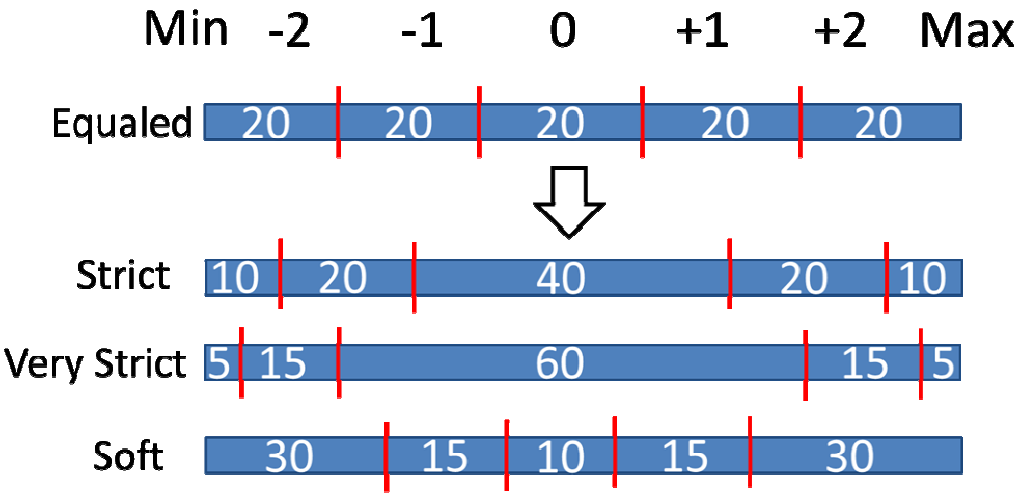


Fig. 7-10 New range of changing rewards

を 15%ずつ,「+2」,「-2」の値域を 5%ずつとったものを考える（Very Strict 案）．また，別の分割法として「0」の値域を 10%,「+1」,「-1」の値域を 15%ずつ,「+2」,「-2」の値域を 30%ずつとったものも用意した（Soft 案）．こちらは“「0」を少なく与え「+2」,「-2」を多く（値域を緩和）して与える”という，教師 C の教え方を模した分割法である．

これら 3 種の新たな値域分割法を用いて，7・4 と同様に客観報酬を 5 段階の離散値に変換し，変換後の報酬に Q 学習を適用する．これにより獲得される行動形態を調べることで，報酬分布が学習結果に与える影響を検証する．

Fig.7-11 及び Fig.7-12 に 4 種類の値域分割法から獲得された行動形態 1 ステップあたりの移動量と，獲得された行動形態の数をまとめたものを示す．Fig.7-11 より，4 つの値域分割法の中で最も優れた行動形態を与えたのは予想外に Soft であった．獲得された行動形態は 1 ステップあたりの移動量が 12.3[mm]に増加し，行動形態の数も 2 パターンまで絞られていた．Strict では移動量が 11.3[mm]と増加したものの Soft には及ばず，行動形態の数が 18 パターンまで増加してしまった．Very Strict では移動量が 6.3[mm]と大きく悪化し，行動形態の数は若干減少したものの，パターン一つ一つが長くなり，非効率的な行動形態が獲得される結果となった．

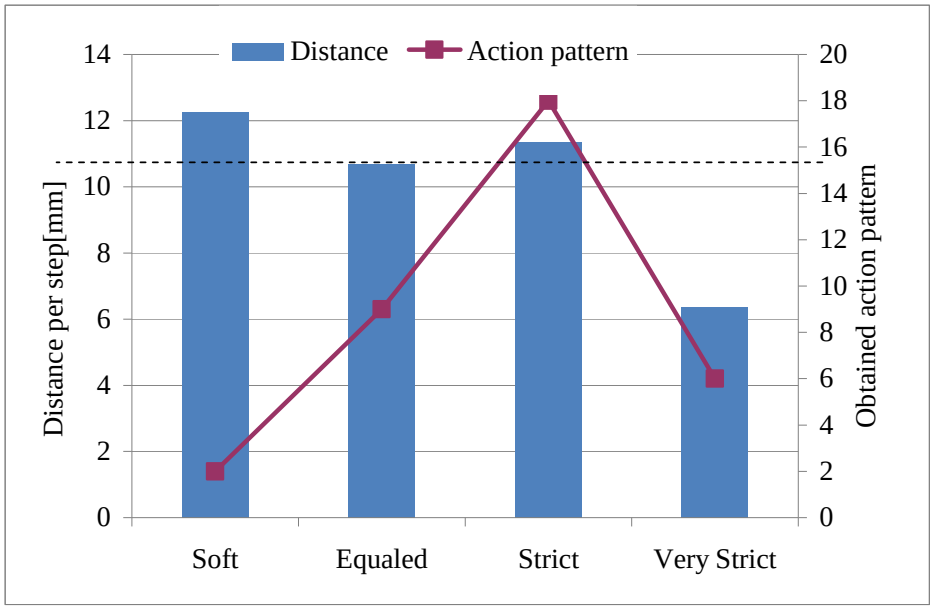


Fig. 7-11 Comparison of the learning results – each changing reward - -objective reward vs. subjective reward

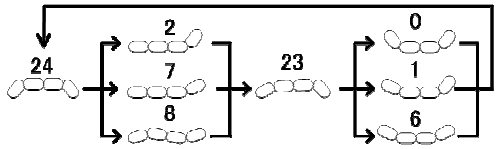
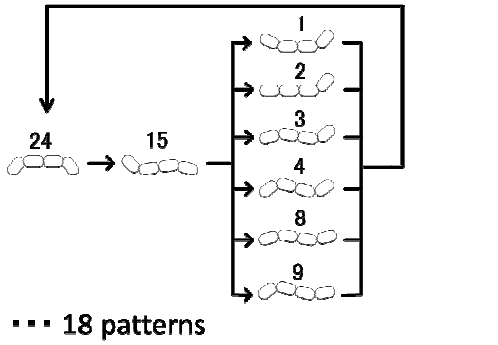
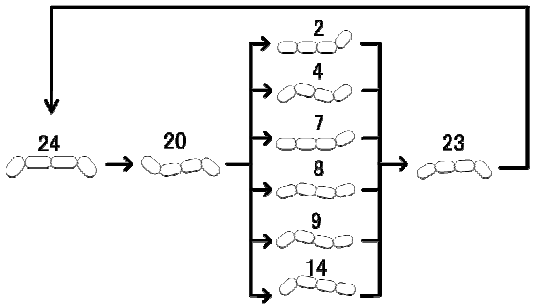
Teacher	Obtained motion forms	Distance per step [mm]
Equaled		10.7
Strict		11.3
Very Strict		6.4
Soft		12.3

Fig. 7-12 Schematic diagram of the learning results - each changing reward -

7･7･3　考察

“「0」を多く与え「+2」,「-2」を厳選して与える”ことが機械よりも優れた人間の評価の特徴で，その特徴を客観報酬の与え方の要素に取り入れることで，学習結果が改善されることを期待した．しかしながら，結果として最も行動形態が改善されたのは予想外に Soft＝“「0」を少なく与え「+2」,「-2」を多く与える”という分割法であり，Strict, Very Strict では Soft ほど行動形態を改善することができなかった．このことは，単純に“「0」を多く与え「+2」,「-2」を厳選して与える”という報酬の比率だけの問題ではないことを意味する．

この原因について考察する．「+2」,「-2」を厳選するという報酬の与え方は，特定の行動を強調して教えることができ，報酬にメリハリをつけるという点から優れた与え方と考えられる．第 5 章 5・4・3 でも，報酬を累乗しメリハリをつけることで，学習速度が向上することを述べた．しかしこの手法の問題点として，$\max\sum r^n$ の最適化問題の解と $\max\sum r$ の解が必ずしも一致しない点があることを述べた．つまり Q 学習は将来にわたって得られる累積的な報酬和を最大化（$\max\sum r^n$）する問題であって，ある一つの行動で得られる報酬を最大化（$\max\sum r$）する問題ではないということである．従って単に現時刻の「＋2」,「-2」の比率を減らして行動を厳選するだけでは不十分で，優れた行動形態の一部分を構成している動作（以後，重要遷移動作）に対して「＋2」を与えられているかどうかがポイントとなる．

このことから考えられることは，優れた教示を行った人間は，ロボットの動きに報酬を与える際に，現時刻のみの判断をしているわけではなく，未来の動作を考慮した価値判断をして報酬を与えていた可能性があるということである．この可能性が事実であれば，この能力は機械には備わってはいない人間特有の能力であると言える．7・6 節にて，人間は過去の動作と比較して相対的な評価を行っているのではないか？との仮説を述べたが，その事実は確かめられなかった．しかし人間は過去ではなく未来の動きに着目しているとすると，それは大変興味深いことである．

次節では人間がどのような行動に「＋2」の報酬を与えたかに注目し，その結果から人間の報酬の与え方の特徴を推察することを試みる．

7･8　人間の優れた教示能力Ⅲ:形状認識

7･8･1　仮説

前節までの結果から，優れた行動形態を獲得させる要因は，現時刻の移動距離だけでなく，優れた行動形態を構成する上で必要な「重要遷移動作」に対して「+2」を与えられるか？ということが要因であると推察される．この可能性を検証するにあたり，客観報酬よりも優れた学習結果を与えた教師 A, D が「+2」報酬を与えた行動に注目した．なお教師 C は＋2 をトータル 42 個の遷移パターンに与えており，この点から行動の優劣を絞り込んでおらず，教師 A,D とは何らか違った観点で評価をしていたと推察され，ここにも教示の主観性が存在することがわかる．

Fig.7-13 に教師 A, D が「+2」報酬を与えた行動を示す．行動遷移パターンは全部で 625 通り（25

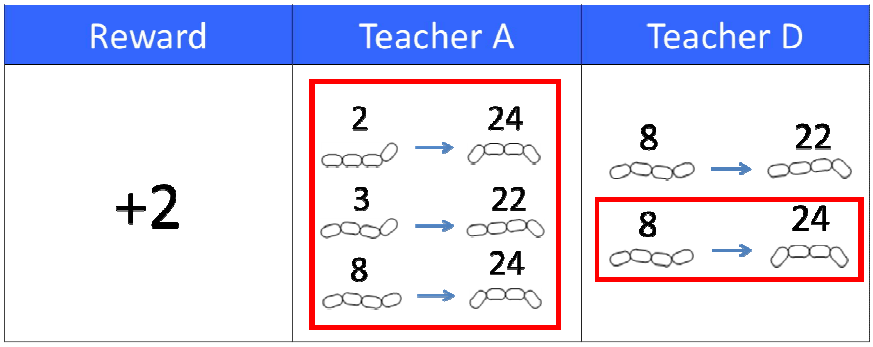


Fig. 7-13 Action that is highly prized by human

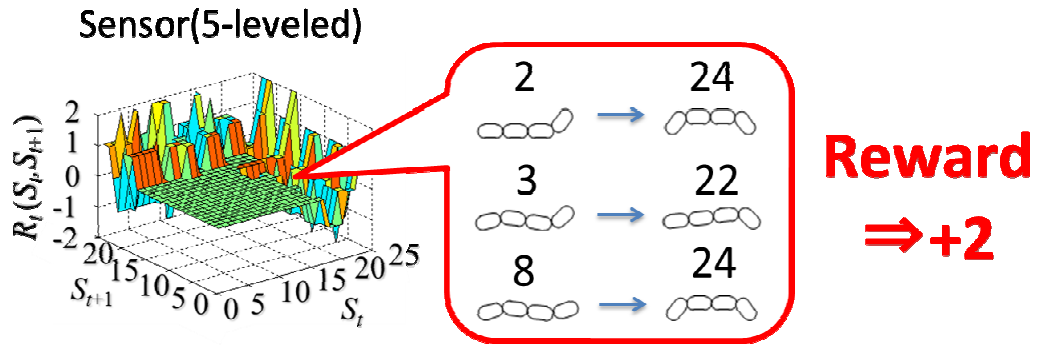


Fig. 7-14 Application teacher A information to Sensor(5-leveled) reward

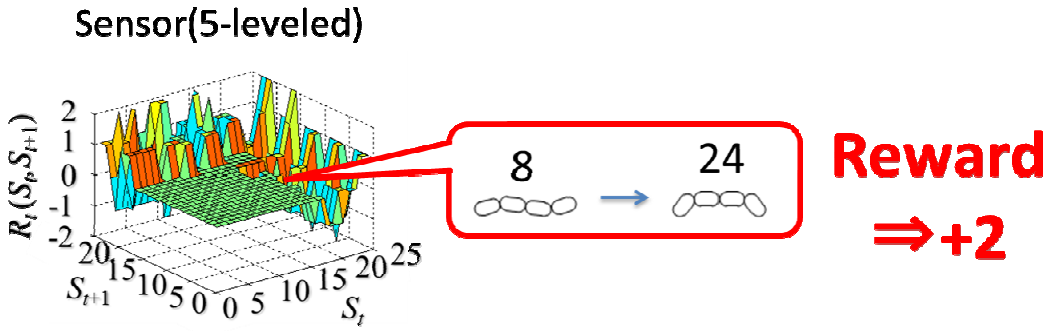


Fig. 7-15 Application teacher D information to Sensor(5-leveled) reward

×25) あるが, 教師 A,D が+2 を与えた行動遷移パターンはわずかに 4 パターンであった. さらに, この中で 2⇒24，3⇒22，8⇒24 は客観報酬（5 値）において「＋2」ではなく，「+1」と判定されている行動であった．つまり，人間はこの 3 パターンの遷移動作については，単純な現時刻の移動量以上に，その行動を評価していた可能性が伺える．

7･8･2　検証

次にこれら特定の行動に「+2」を与えることが，学習結果にどのような影響を与えるか検証する．ここでは 7・4 で 5 値化された客観報酬に，教師 A,D が高く評価した行動の知識を適用することを考える．教師 A, D の知識を適用する場合の模式図を Fig.7-14 及び Fig.7-15 に示す.

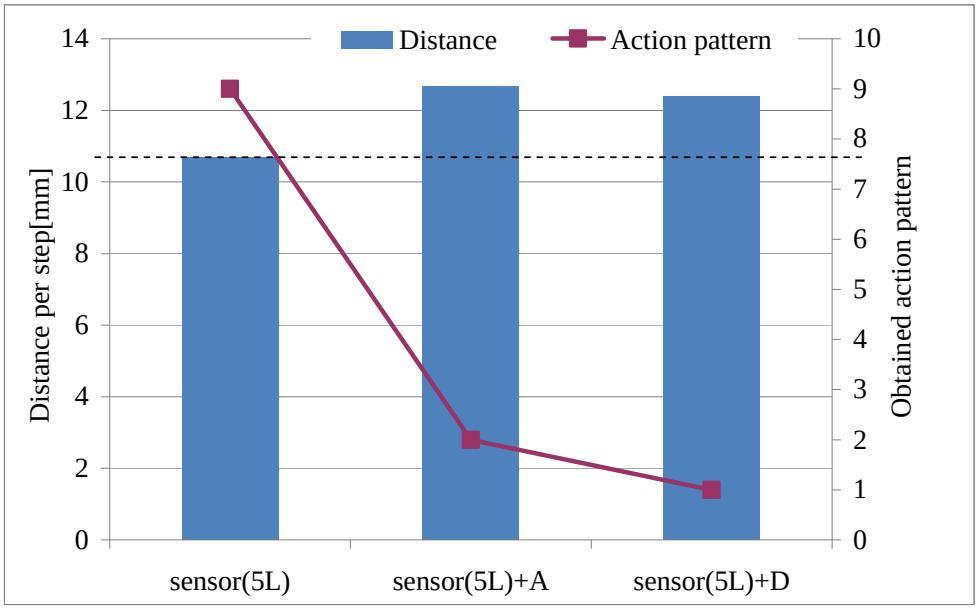


Fig. 7-16 Comparison of the learning results
- application of teacher information -

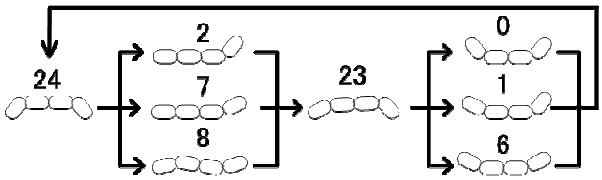
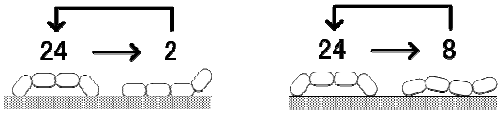
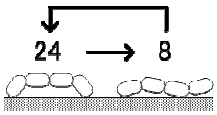
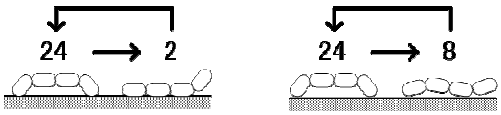
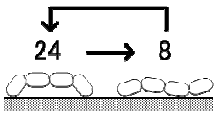
Teacher	Obtained motion forms	Distance per step [mm]
Sensor (5-leveled)		10.7
Sensor (5-leveled) + A		12.7
Sensor (5-leveled) + D		12.4
Teacher A		12.7
Teacher D		12.4

Fig. 7-17 Schematic diagram of the learning results
- application of teacher information -

センサによる客観報酬（5 値）をベースとし，教師が「+2」を与えた行動の報酬を，客観報酬（5 値）でも「+2」に書き換える．教師 A の知識を適用するのであれば 2⇒24，3⇒22，8⇒24 の 3 つを，センサで判断された「+1」から「+2」に書き換える．教師 D の場合は 8⇒22 が客観報酬でも「+2」と判断されているため，実質 8⇒24 のみを「+1」から「+2」に書き換えることで，教師の知識を客観報酬に適用したものとする．それ以外の報酬については操作しない．教師が高く評価した行動の知識を適用した後，Q 学習を適用し，行動形態を獲得させる．

知識適用前：sensor(5L)，A 知識適用：sensor(5L)+A，D 知識適用：sensor(5L)+D として，獲得された行動形態 1 ステップあたりの移動量と，獲得された行動形態の数をまとめたものを Fig.7-16 に示す．また，獲得された行動形態の模式図を Fig.7-17 に示す．

Fig.7-16，Fig.7-17 より，教師 A, D が高く評価した行動を「+2」に書き換えたことで，獲得される行動形態が大きく改善されていることがわかる．どちらの知識を適用した場合でも，移動量が増加しただけでなく，獲得される行動形態が各教師の報酬とまったく同じものになっていることがわかる．

7･8･3　考察

7・8・2 より，「重要遷移動作」に「+2」を与えることが優れた行動形態を与える要因ではないか？という推察は正しかったと考えられる．わずか 1〜3 個の行動の報酬を「+1」から「+2」に置き換えただけで，獲得される行動形態が大きく改善されるのは興味深い事実である．移動量が増加しただけでなく，獲得される行動形態が教師と全く同じものになったということは，置き換えた「重要遷移動作」への報酬が，教師の行動形態の獲得にも大きな影響を与えているはずである．

それでは，優れた行動形態を獲得させる「重要遷移動作」とはどのような行動なのだろうか？先にも述べたように，これは単純な移動量だけでは説明がつかないと考えられる．次節ではこの「重要遷移動作」について，教師 A, D が高く評価した行動を手掛かりに調査を進めていく．

7･8･4　形状認識能力

7･8･1 に示したように, 教師 A, D は合計 4 つの行動に対して「+2」報酬を与えていた（Fig.7-13）．これら 4 つの行動に共通する特徴がないかを調べる．ここで Fig.7-18 より，行動前の状態と行動後の状態はそれぞれ似通った形のものが選ばれているのではないかと考え，この「形状」に注目して調査を行った．4 つの行動に含まれる 5 つの状態について，前部（＝頭部）モータ，後部（尾部）モータの角度をまとめたものが Fig.7-19 である．まず行動前の状態について，頭部モータ角が +52°or +26°，尾部モータ角が 0°or -26°という，頭を上げて尻尾を少し下げる形（ジグザグ，または腹這い）が選ばれていた．一方行動後の状態については，頭部モータ角が+52°，尾部モータ角が 0°or +52°という，頭を大きく下げたブリッジ型の状態が選ばれていた．このように，教師 A, D が「+2」を与えた行動は，含まれる状態の形に共通点が見られた．

重要遷移パターンに「+2」報酬を与えることで学習結果が大きく改善されることは 7・8・2 で示した．本章ではそれを拡張し，人間が「+2」を与えた行動と類似した形状を持つ行動すべてに

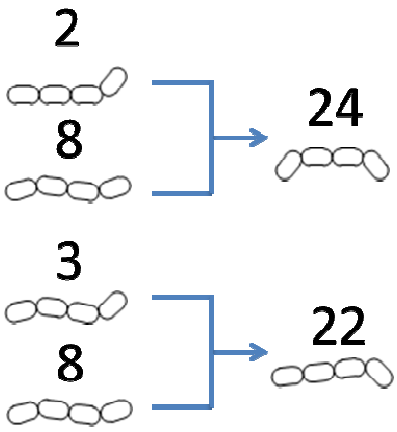


Fig. 7-18 Action which is highly prized by human

State	Head angle[degree]	Tail angle[degree]
2	+52	0
3	+52	-26
8	+26	-26

State	Head angle[degree]	Tail angle[degree]
22	-52	0
24	-52	-52

Fig. 7-19 Details of specific states which is highly prized by human

「+2」報酬を与えることを考える．これにより，「行動形状に注目して報酬を与える」=「良好な学習結果が得られる」という因果関係が存在するのかを検証する．「行動形状」の情報を適用する際は 7・8・2 の手法をベースとし，客観報酬（5 値）の報酬データベースのうち，人間が注目した行動の形に類似しているパターンの報酬を「+2」に置き換えた．具体的には，行動前が状態 2, 3, 7, 8 であり，行動後が状態 22, 23, 24 となるような行動（4×3＝12 通り）を選択し，客観報酬（5 値）におけるこれらの行動の報酬値をすべて「+2」に置き換えた（Fig.7-20）．この報酬に Q 学習

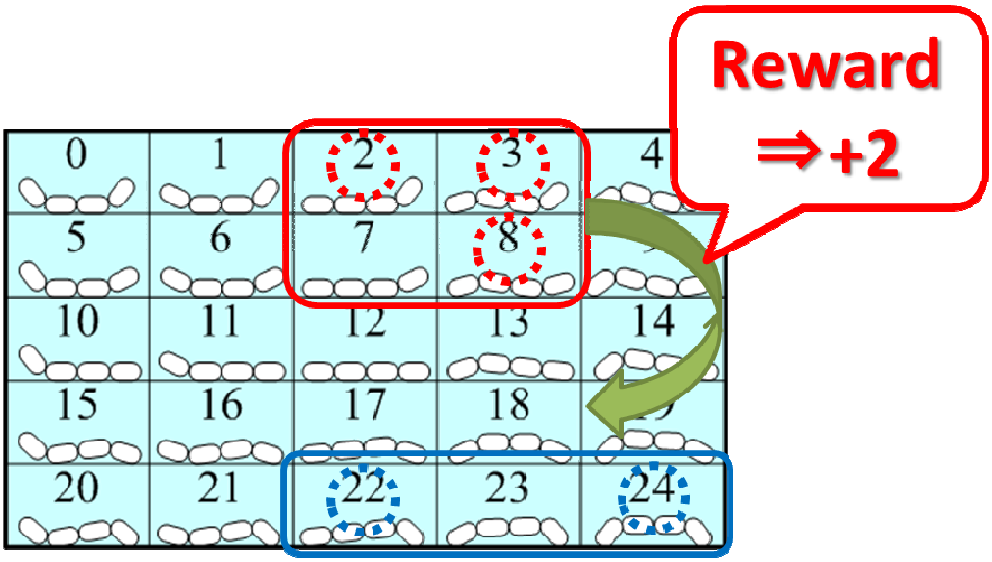


Fig. 7-20 Similar state forms

を適用し，行動形態を獲得させる．

Fig.7-21 に獲得された行動形態 1 ステップあたりの移動量と，獲得された行動形態の数を，Fig.7-22 に獲得された行動形態の模式図を示す．両図は比較のため，「行動形状」を取り入れる前の学習結果も併せて掲載している．Fig.7-21 を見ると，「行動形状」を評価に取り入れた結果，行動形態の移動量が改善され，行動形態の数も減少していることがわかる．また Fig.7-22 を見ると，獲得される行動形態一つ一つのループが短くなり，教師 A, D に近い形の行動形態が獲得されていることがわかる．

以上の結果から，優れた学習結果を与えた人間（A, D）は移動量に加えて「行動形状」を考慮して「+2」報酬を与えており，それが優れた行動形態を獲得させる要因となったことが示された．すなわち，センサ（機械）が行動を現時刻の「移動量」のみで評価するのに対し，優れた教示を行った人間は「移動量＋行動形状」で評価していたと考えられる．

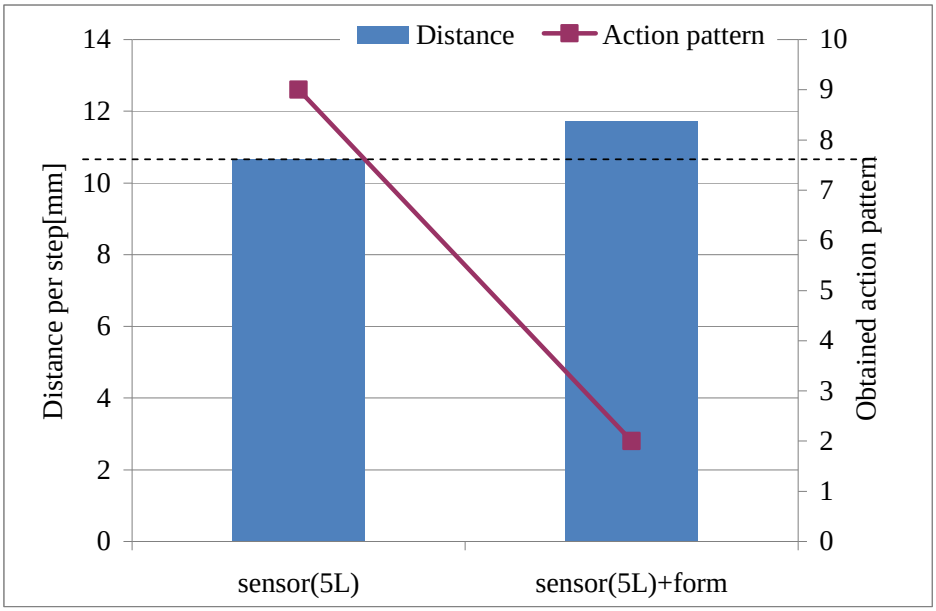


Fig. 7-21 Comparison of the learning results
- application of state form information -

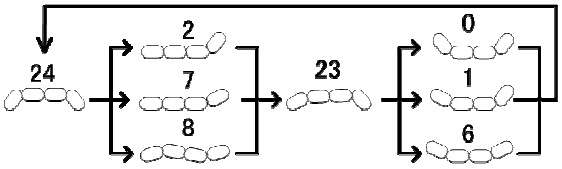
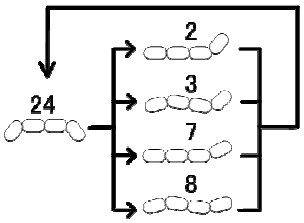
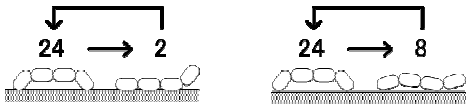
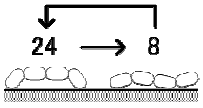
Teacher	Obtained motion forms	Distance per step [mm]
Sensor (5-leveled)		10.7
Sensor(5-leveled) + form information		11.7
Teacher A		12.7
Teacher D		12.4

Fig. 7-22 Schematic diagram of the learning results
- application of state form information -

報酬の分解能が高い場合は，正しい距離情報を緻密に与えられるセンサ報酬の方が優れた学習結果を得られるが，分解能が同一の場合には移動量による評価のみでは行動の優劣が付けられない為，行動形態を絞り込めない．一方人間は移動量に加えて「行動形状」まで考慮して評価するので，行動形態をある程度絞り込めて，結果として優れた前進行動を獲得させることができたものと考えられる．

それではなぜ「行動形状」を考慮するとセンサよりも良好な学習結果が得られるのだろうか？Fig.7-23 に各状態番号から遷移した場合の平均移動距離を示す．例えば ID13 であれば，

13→0, 13→1, 13→2・・・13→24 のそれぞれの移動距離を平均化した値である．

Fig.7-23 を見ると, ID24 からの状態遷移は平均移動距離が最も大きいことがわかる. このため，ID24 へ遷移すればその後に大きな移動が得られると見込めるので，状態遷移の中でも特に ID24 への遷移に高い報酬を与えることができるかどうかが，本タスクでロボットに効率的な前進行動を獲得させるキーファクターとなっている．

Fig.7-24 には，各 ID への遷移に対してセンサ及び人間が与えた報酬の平均値を示している．センサが与えた客観報酬を見ると ID24 への遷移に対して，全 ID の中で相対的に大きな報酬を与えていないことが見て取れる．むしろ報酬値としては小さな値となっている．このことは ID24 に遷移すること自体はそれほど大きな前進距離を得られないことを意味している．一方，優れた学習結果を与えた教師 A,D の結果を見ると，ID24 への遷移に対して，相対的に高い報酬を与えているのがわかる．これは，人間が報酬を与える時点で未来の状態遷移を考慮して価値判断し，それを報酬へ投影しており，これは Q 学習のプロセスそのものを無意識的に行っていることを意味する．機械（センサ）の与える報酬には未来を予見する能力は備わっていない為，ID24 へ移動すれば良好な前進行動を得られることは，報酬を与える時点では判断できず，Q 学習が終了した後に行動価値 Q を観察することで，結果として知ることになる．

報酬の分解能が制限されると，行動価値 Q の分解能も低下する．この為，センサによる客観報酬では連続値→5 値へ報酬の分解能が低下すると，行動価値 Q の分解能も低下する．しかし人間は距離に加えて行動形状も加味して報酬を与えている為, 報酬の分解能が 5 段階距離＋α となる．このことにより，センサによる客観報酬よりも報酬分解能が相対的に高くなり，結果的に得られる行動価値 Q の分解能も高くなり，良好な学習結果が得られたと考えられる．

この行動形状を意識した，すなわちロボットの未来の動作を考慮した報酬の与え方は，機械にはない人間の優位性であるといえる．しかし興味深いことに，これは全ての人間で共通して行っているわけではなく，ID24 への遷移に大きな報酬を与えていない被験者もおり，ここにも教示の主観性が存在することが見て取れた．

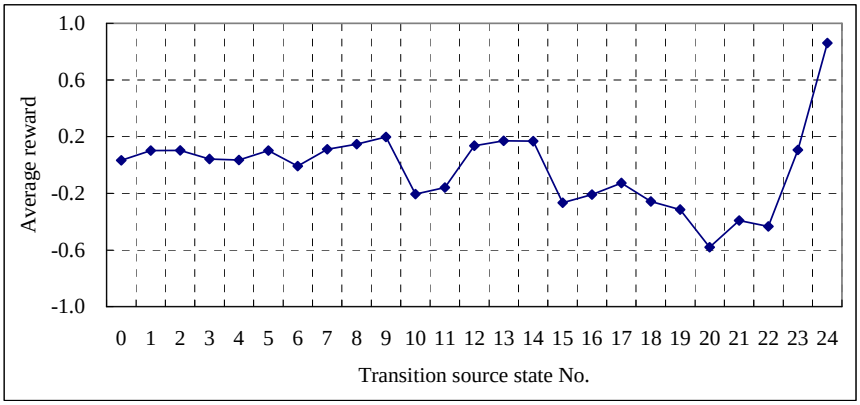


Fig. 7-23 Average distance moved from each state

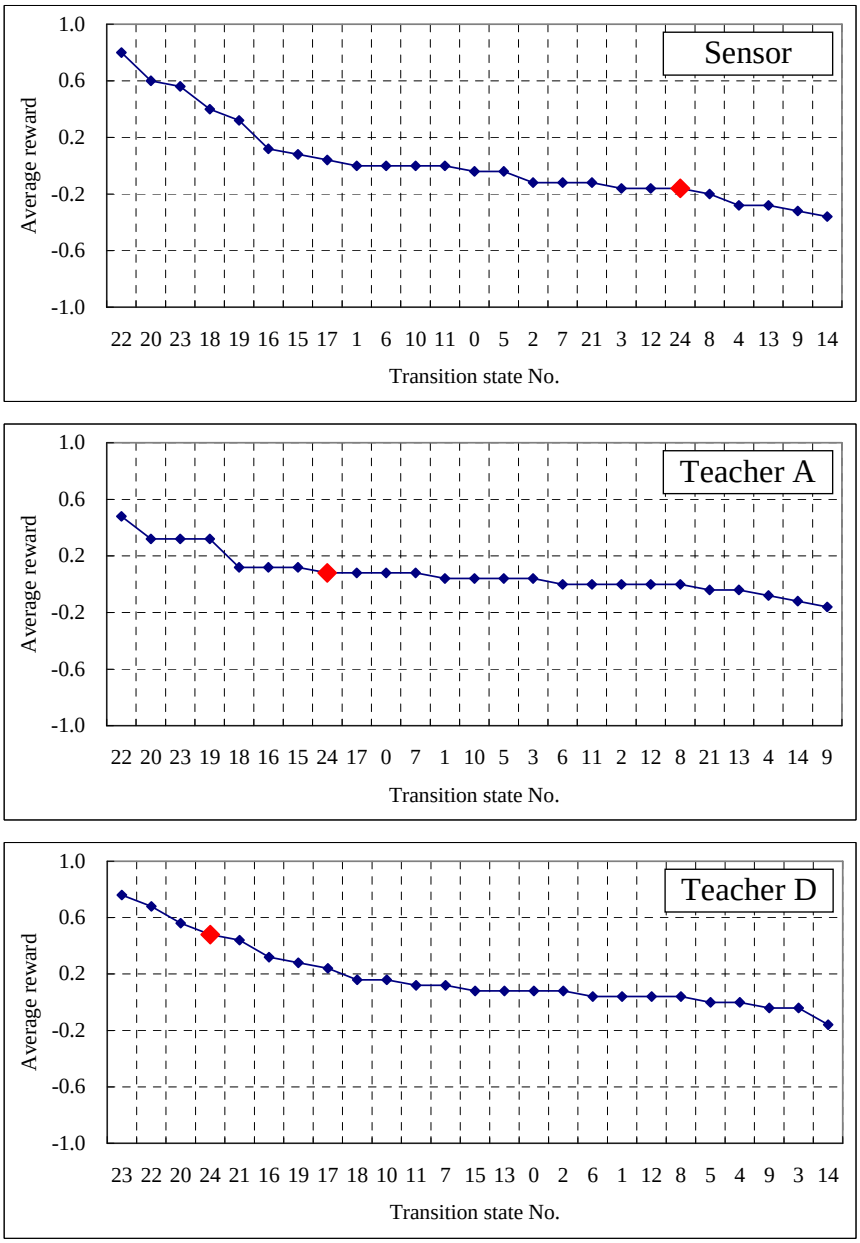


Fig. 7-24 Average distance moved from each state

7-9　機械学習における投影教示

最後に，人間（A, D）はなぜ Fig.7-18 で示した行動形状に注目し大きな報酬を与えたのだろうか？またその行動形状にはどのような意味があるのだろうか？ここからは推測になるが，現実の生物の動作パターンを想起したのではないか？と考えている．本研究ではイモムシ型ロボットを用いたが，今回獲得された行動形態は現実の生物の「シャクトリムシ」に近い動きである（Fig.7-25）．「+2」を与えられた状態遷移は，シャクトリムシが体を曲げて前進のためのエネルギーを蓄える動きに相当する．この動きはセンサから見ればエネルギーを蓄えているという認識は報酬を与える時点では判断できない為，大きな価値を置くことは無い．しかし人間はこのエネルギーを蓄える動作が次につながる重要な動作であると経験的に判断し，この動作に価値を置き報酬を与える時点で反映させることができたと考えられる．

このように人間はこれまでに自身が見たり経験したり感じたことを投影して，ロボットの評価を行っていたのではないか？と推察される．この「経験の投影」は機械には無い人間の特徴であり，客観報酬よりも主観報酬が優位性を与える要因であったと考えられる．

本実験のロボットは現実の生物を厳密に再現したものではない為，この仮説を検証する上では今後対象ロボットに工夫を加えたい．例えば現実に存在する形態の生物型ロボットと，存在しない形態のロボットとで，人間が与える報酬や学習結果にどのような差が生じるのか？を検証することにより，「経験の投影」がなされているのかを検証することができる．また対象を人型の歩行ロボットにすることで，自分自身が日頃行っている歩行動作の投影，すなわち自己投影に基づく教示がなされるのかを検証できると考えられる．

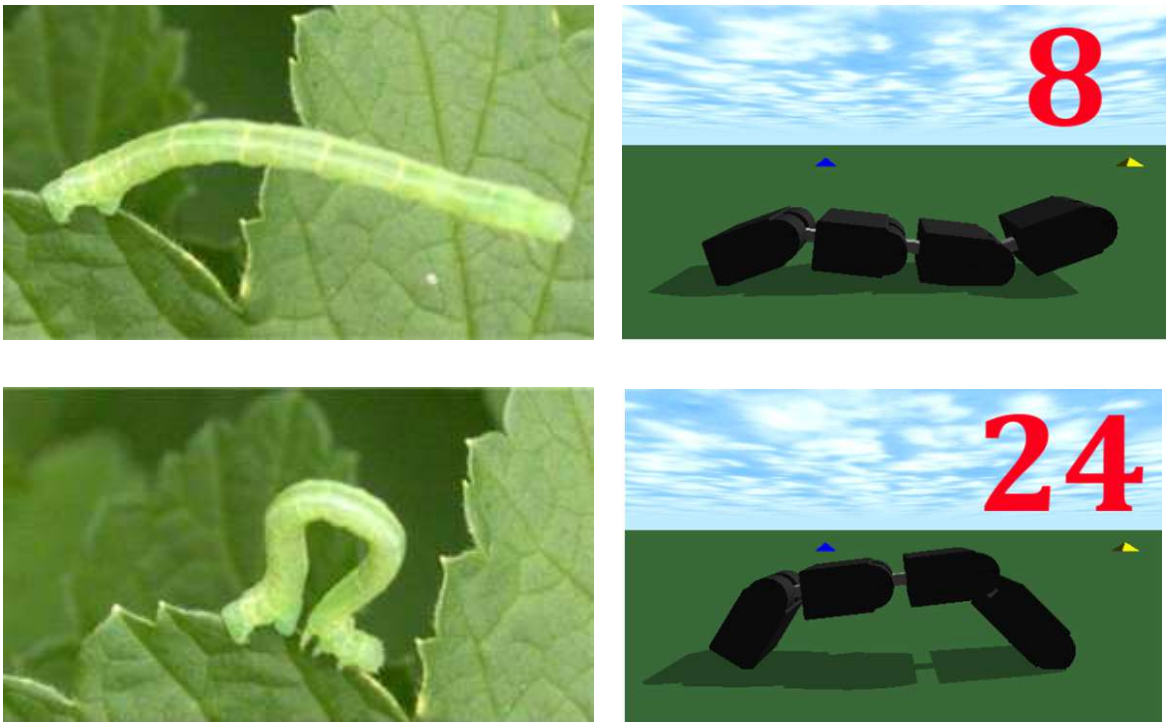


Fig. 7-25 Looper caterpillar and caterpillar robot

また主観報酬の研究を今後継続的に進める上では，評価方法や評価指標についても工夫を加えたい．今回は 5 名の被験者に「ロボットを前に進ませたい」という目的のみを伝え，行動の良し悪しを 5 段階で評価してもらったが，例えば被験者に「最も良い行動形態を一つ獲得させたい」という目的を伝えていたら，教示方法や学習結果が異なっていたことが予想される．学習目的を変えると，学習結果がどのように変化するかは興味深い問題であり，今後検討を進めたい．また本研究では評価指標を「移動距離」という客観的指標での学習結果を論じてきたが，人間の主観報酬がより大きな効果を発揮するのは，動きの「しなやかさ」や「優雅さ」といった主観的指標を最大化する学習であると考えられる．このように，今後は本研究の発展系として，評方法や評価指標の拡張も検討し，教示の主観性が与える学習への影響や，人間の優れた能力を更に明らかにしていきたい．

7-10 まとめ

本章では主観報酬に基づく強化学習について検討を進め，主観報酬の与え方と学習結果を考察することで，機械には無い人間の優れた教示能力を明らかにした．

本検討から明らかになった人間と機械の教示の違いは，移動量のみに着目して報酬を与える機械（センサ）に対して，人間は移動量＋「行動形状」でロボットの行動を評価しており，この違いが，報酬の分解能が制限された中でも，行動の優劣を教示する有効な手段となっていることを述べた．また，これらの行動形状を意識した評価は，人間がこれまでに経験した知識を投影する投影教示の可能性があることを示した．

第8章
結論

本論文では，機械学習における人間と機械の協調学習に焦点を当て，その研究過程を通じて人間の優れた能力，機械の優れた能力の一端を明らかにした．

本論文より得られた主な結論は以下の通りである．

(1)　人間と機械の優れた能力の理解，及び協調学習の枠組みの提案

フレキシブルアームロボットの制振問題をタスクとし，ニューラルネットワークを適用し，振動を抑制するコントローラを機械に学習させた．人間が得意とする優れた能力は「タスクの大枠を判断する能力」であることを示し，この能力を生かして機械の探索空間を狭めることで，機械の得意とする最適化問題処理能力を最大限生かす協調学習の枠組みを提案した．具体的にはモデル/コントローラ繰り返し学習法を提案し，機械が自らモデル，コントローラを更新する自律性の高い学習システムを構築し，この学習システムに対して人間が必要最小限の教示を行うことで，人間と機械の得意・不得意をカバーし合う協調学習が可能となることを示した．

(2)　人間の報酬操作による機械学習支援

イモムシ型ロボットの前進行動獲得をタスクとし，強化学習を適用した．はじめに距離センサ情報に基づくシンプルな教示（報酬）のみで，ロボットが前進行動を獲得できることを示した．次に報酬を人間が操作することで，ロボットの学習が促されることを示した．具体的には，学習初期に On-Off 型のフィルタ報酬を与えることで，最終収束行動価値に近い推定値かつ，ばらつきを持った行動価値を学習初期に配置することができ，より効率的に学習が進むことについて述べた．本手法は人間の学習過程に倣い，学習初期には行動の良し悪しを大まかに教示し，学習の深度に応じて詳細に教示を行ったものであり，(1) にも示したような人間が大枠を教示して機械の探索空間を狭め，機械の得意とする最適化処理をサポートする手法である．

次に報酬値を累乗操作することにより，有益な報酬が強調され，学習の収束性を向上させることができることを述べた．ただし本手法は報酬を非線形的に変化させている為，システムの報酬情報を壊し所望の学習結果が得られない可能性があり，最適なべき数に関しては検討の余地を残した．Q-Learning は一連の行動の報酬を最大化することであり，本手法の $\max \sum r^n$ の最適化問題の解は $\max \sum r$ の解と必ずしも一致しないことが本手法の問題点となる．

(3) 教示の新たな視点（教示の主観性・客観性）の提起

機械学習における教示の新たな視点として，教示の主観性・客観性の問題を提起した．**「自分ひとりの考え方や感じ方」**である**「主観性」**を強化学習の中の報酬に取り入れ，機械学習への影響を検討した．はじめに距離センサが連続値に与える客観報酬と，人間が 5 段階で与える主観報酬を用い Q 学習を適用した．獲得された行動形態を比較すると，客観報酬の方が優れた行動形態を獲得させる結果となった．これはタスクの評価関数が「距離」（客観指標）である為，正確に距離情報を教示する客観報酬での学習結果に優位性がある当然の結果とも言えた．しかし，客観報酬と主観報酬の分解能を同一の 5 段階とした場合は，主観報酬の方が優れた行動形態を獲得させることを明らかにした．この事実から，人間が与えた主観的な報酬には，センサにはない，機械の学習をサポートする有益な情報が含まれているのではないか？という仮説を立てると共に，教示の主観性という新たな問題提起を行った．

(4) 教示の主観性と機械学習への影響

主観報酬の学習優位性について，人間が与えた報酬値に基づき調査した．はじめに優れた学習結果を与えた人間は，過去の動きと比較して相対的に現在の動きを評価する「相対教示」を行っているのではないか？という仮説を立てた．検証の結果，人間の多くが相対教示を行っていることが確認されたが，それが必ずしも優れた前進行動獲得につながる有効な教示というわけではないことを明らかにした．

次に人間の与えた報酬の分布に注目した．報酬の分布は被験者により大きく異なっており，主観性の強い報酬であった．優れた行動形態を獲得させた教師は，客観報酬と比較して「0」を多く与え，「+2」，「-2」を少なく与えていたことがわかり，このメリハリのついた報酬分布の特徴をモデル化して客観報酬に取り入れ，学習を行った．しかし，この報酬分布の特徴を取り入れただけでは、よりよい行動形態に改善することはできないことを示した．

最後にロボットの行動形状に着目して解析を進めた．検証の結果，優れた教示を行った教師が大きな報酬（＋2）を与えた行動には，移動距離ではなく行動形状に類似性があることがわかった．そしてこの行動形状を報酬に取り入れることで，移動量のみの報酬よりも，学習結果が向上することを示した．良好な学習結果を与えた人間は，現在の移動量のみを評価しているのではなく，その行動が次につながる良い行動なのか，すなわち未来の動作への影響を行動形状から無意識的に判断していることが示された．

なぜ人間はこのような未来を見据えた教示を行っていたか？について，投影教示の可能性を論じた．人間が大きな報酬（＋2）を与えた動作を観察すると，現実に存在する生物（シャクトリムシ）の動きと類似していることがわかり，人間はこれまでに自身が見たり経験したり感じたことを投影して，ロボットに教示を行っている可能性を示した．興味深いことに，被験者によっては，このような未来を見据えた教示を行っていないものもおり，ここに教示における主観性と学習の因果関係の興味深い問題が潜んでいることが確認された．

第9章
今後の展開
～機械学習における主観性の問題～

本論文では，機械学習における人間と機械の協調学習に焦点を当て，その研究過程を通じて人間の優れた能力，機械の優れた能力の一端を明らかにした．

人間は経験的にタスクの大枠を判断する能力や，直感的に物事を判断する能力に長けていると言われている．これらの能力は，コンピュータでの表現が難しい「自分ひとりの考え方や感じ方」を扱った「主観性」の問題である．

本研究では機械学習に主観性の問題を取り込み，教示の主観性が機械の学習に各種影響を与えることの一端を示した．

機械学習の大きな目的は元来，機械に自律的な動作・行動獲得を求めることにある．この観点での機械学習の活用は，設計行為の自動化や合理的な戦略獲得といった，個人に依存しない最適な結果・結論を求める活用方法と言える．このような問題は，誰にとっても望ましいことが一致している問題であり，評価指標が客観的な指標である．つまり客観指標の最大化問題としての機械学習活用であり，個性を排除することを念頭に置いている．(Objective Learning)

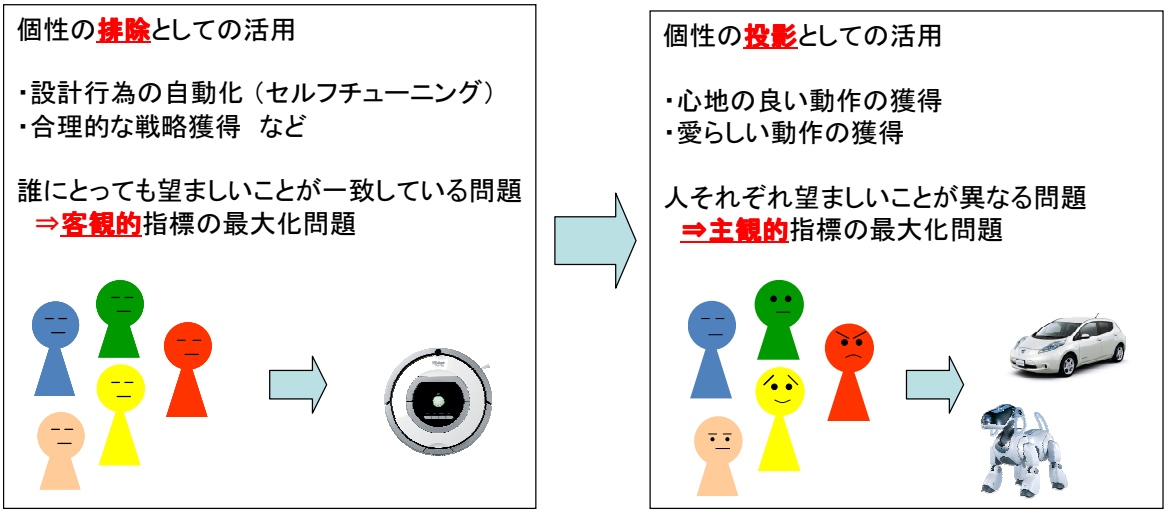


Fig. 9-1 Shift to Subjective Learning

一方，物事の良し悪しが，人それぞれ考え方や感じ方により異なる問題もある．例えば心地の良い動作の獲得や，愛らしい動作の獲得と言った問題である．これらは評価指標が主観指標であるため，主観指標を最大化する機械学習活用と言える．ここでは個性を排除するのではなく，個性を投影するツールとして機械学習を活用することになる．**(Subjective Learning)**

主観指標の最大化問題に対するアプローチは，例えば「心地よさ」や「愛らしさ」を何らかの数値で定量化し，客観的な表現に置き換えた上で機械に教示する方法が考えられる．しかし，人間の感性や価値観を定量的に表現することは極めて困難な問題である．従って，主観的な教示と，主観的な動作獲得の間をダイレクトに結ぶ手段として，機械学習を活用できれば，様々な分野へ応用することが期待できる．

本研究において，第 2 章，3 章では，ニューラルネットワークを用いたアームの制振問題をタスクとした．この問題では「振動を素早く止める」という極めて客観性の高い問題設定に対して機械学習を用いた．すなわち「振動を素早く止める」という評価関数は客観的であり，誰もが共通の目的と結果を共有できるタスクである．一方例えば「しなやかに止める」「滑らかに止める」といったタスクであればどうか？これらは人それぞれ「しなやかさ」や「滑らかさ」の判断基準が異なるから，主観的なタスクと言える．

第 6 章，7 章では，強化学習を用いたロボットの前進行動獲得をタスクとした．ここでも「前進距離の大きな行動を獲得する」という評価関数は客観的であった．この問題も「優雅に前進させる」「躍動感をもって前進させる」といったタスクであれば，それは主観的な問題となる．

このように機械学習に主観性の観点を組み込むことで，機械へ人間の個性を投影することが可能になると考えられ，今後この分野の検討が進めば，このコンセプトは様々なアプリケーションへの展開が期待できる．本研究では主観を取り込んだ機械学習の一端を明らかにしたが，今後この分野の研究を進めて，発展させていくには，以下の検討が必要となる．

(1) 主観評価方法の検討

本研究では 5 名の被験者に対して検証を行った．主観性を検証する上では，統計的処理が欠かせない．この為，必要十分な被験者数の確保や，特徴量データの統計的処理を適切に行う必要がある．また性別，年齢，文化の違う被験者に協力してもらうことで，より興味深い知見が得られるであろう．今後，認知心理学の分野の知見も生かし進めていきたい．

(2) 対象ロボット（システム）の選定

個性の投影としての活用を検討する上では，対象ロボットの選定も重要となる．人間が日頃目にしている実在の生物モデルと，見たこともない実在しない生物モデルとで教示にどのような差が表れるのか？あるいは人型のロボットを用いた場合には，自身が持つ構造を意識した自己投影が行われるのかなども興味深い．ロボティクス分野，生体工学分野との連携を進めたい．

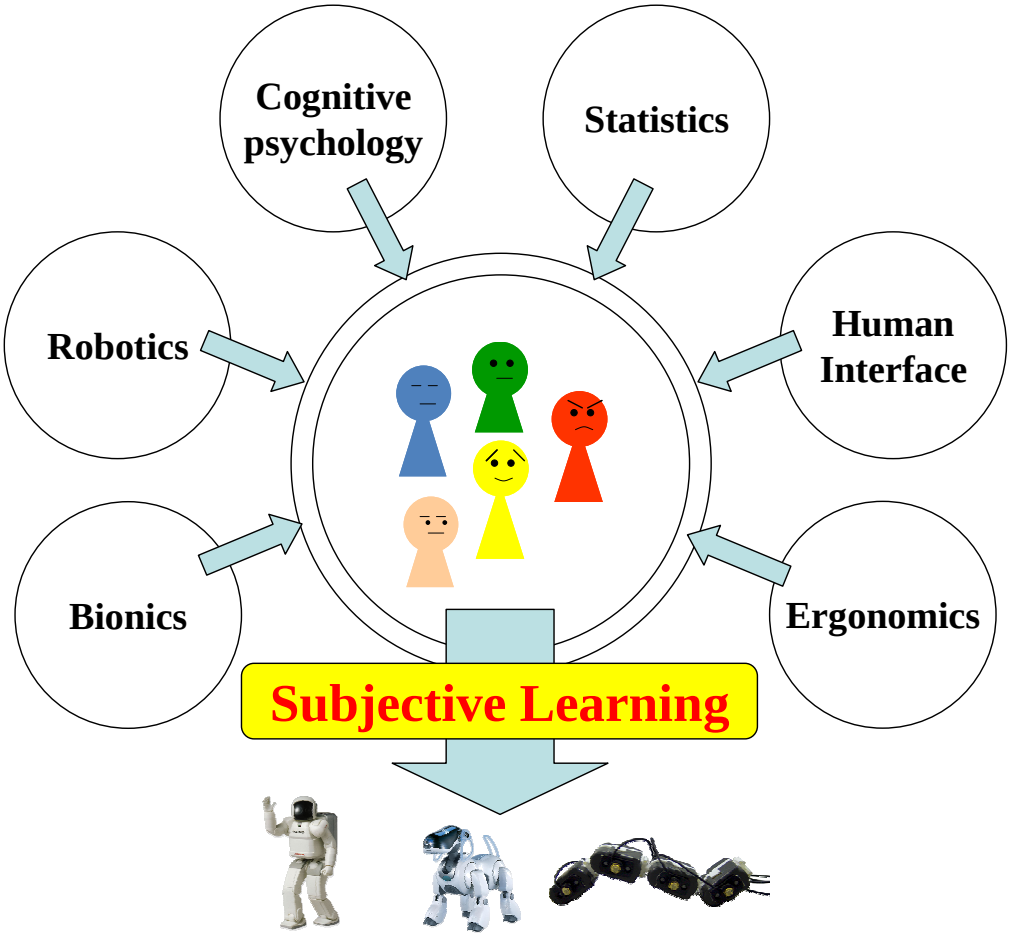


Fig. 9-2 Associated research of Subjective Learning

(3) 5 感を活用した教示の検討

教示を行う上で，対象ロボットの動作に対して，人間の視覚情報に制限を加えたり，聴覚や触覚情報を利用した教示を行うなどし，5 感情報と学習結果との関係を検討することで，興味深い結果が得られると考えられる．また心地よさや快適さを投影する上では欠かせない人間工学やヒューマンインターフェイス分野との関連も意識したい．

Subjective Learning の今後の展望（Fig.9-3）

ハードウェア（プロダクト）の観点から言えば，近年，大量生産型の商品提供から，個人の感性や価値観に基づいた少量生産型の商品提供が増えてきている．ファッションなど元来，主観性の求められる分野はもちろんだが，靴の履き心地やベッドの寝心地といった，人それぞれにあった機能性が求められる分野や，香り，肌触りといった癒し・娯楽分野へと，少量生産型のプロダクト提供は広がりを見せている．また 3D プリンタの登場などにより，今後少量生産に対するコスト的・時間的制約は徐々に取り除かれていくだろう．

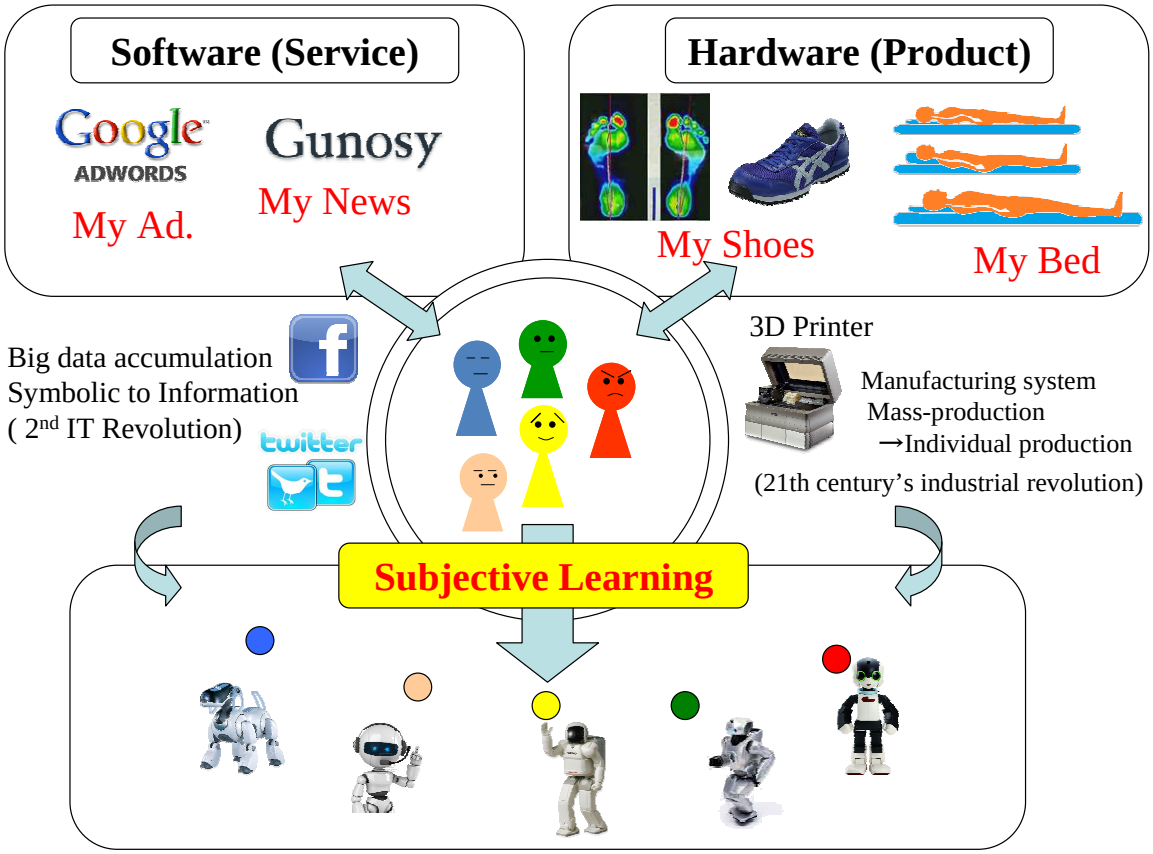


Fig. 9-3 Framework of Subjective Learning

ソフトウェア（サービス）の観点から言えば，新聞やテレビといった大衆向けの情報を提供するサービスから，個々人にあった広告，ニュース，リコメンド機能などが爆発的に増加している．またツイッターや Facebook といった個人の感情・価値観を受発信する仕組みや，それらの大規模データを活用するビッグデータ処理の研究が盛り上がりを見せるなど，ソフトウェアにおいても，個人・個性をターゲットとしたサービス提供が広がっている．

このように 21 世紀は，大量型 ⇒ 少量型，大衆型 ⇒ 個人型，グローバル型⇒ローカル型といった，個人の価値観にあったサービス展開が求められる時代と言える．

機械学習もまた，客観的教示に基づく，個性の排除としての機械学習（**Objective Learning**）から，主観的教示に基づく，個性の投影としての機械学習（**Subjective Learning**）へとシフトすることで，機械学習の新たな応用展開が期待できると考える．

序章で，機械は人間を超えられるか？という哲学的問いかけに対して No.と答えた．これは評価関数自体を機械は学習することが出来ないので，評価関数を設計する上位の人間を，概念的に越えることは出来ない，という意味であった．しかし，主観的な評価関数に基づき学習を行った機械に，教示者特有の何らかの個性が投影されるとするならば，その個性を持ったロボット同士が協調することで，新たな個性や価値観なるものが生まれるかもしれない．そしてそれは，最初に教示を行った人間の想像を超えたものであるかもしれない．

謝辞

本研究を進めるに当たり，懇切丁寧なご指導を賜った横浜国立大学大学院 藪田哲郎 教授に深く感謝の意を表し，厚く御礼を申し上げます.

機会ある度に貴重な助言を頂いた豊田希 研究教員，東京大学 原正之助教，日本精工株式会社林俊樹氏，にも厚く御礼申し上げます.

また，本研究を共に遂行した黒田将史氏をはじめ，実験にご協力頂いた藪田研究室に在籍した諸氏に厚く感謝の意を表します.

また，貴重なお時間を頂き博士論文を審査して頂く横浜国立大学大学院 藪田哲郎 教授，高田一 教授，眞田一志 教授，佐藤恭一 准教授，前田雄介 准教授に深く感謝の意を表し，厚く御礼を申し上げます.

文献

＜学習理論一般＞

ＮＮ

[1] WILLIAMSON Matthew M, “Neural control of rhythmic arm movements”, *Neural networks : the official journal of the International Neural Network Society*, Vol. 11, No. 7, (1998), pp.1379-1394.

[2] WATANABE S, “Ultrasonic robot eyes using neural networks”, *IEEE Trans. of UFFC*, Vol. 37, No. 2, (1990), pp.141-147.

[3] MIYAMOTO Hiroyuki , KAWATO Mitsuo, “A tennis serve and upswing learning robot based on bi-directional theory”, *Neural networks : the official journal of the International Neural Network Society*, Vol. 11, No. 7, (1998), pp.1331-1344.

[4] 米川雅人，黒川弘章，“Pulse Coupled Neural Network を用いた類似画像検索”， 電子情報通信学会技術研究報告. NLP，非線形問題，Vol. 111, No. 498, (2012), pp.57-62.

[5] T. Tsuji, K. Ito, M. Pietro, “Neural Network Learning of Robot Arm Impedance in Operational Space”, *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, Vol. 26, No. 2, (1996), pp.290-298.

[6] 小平実，大友照彦，田中敦，岩月正見，大内隆夫，“ニューラルネットを用いた移動ロボット車の障害物回避走行制御”，電子情報通信学会論文誌. D-II，情報・システム，II-情報処理， J79-D-2(1), (1996), pp.91-100.

[7] SANGER T. D, “Neural Network Learning Control of Robot Manipulators Using Gradually Increasing Task Difficulty”, *IEEE Trans. Robot. Auto*, Vol. 10, No. 3, (1993), pp.323-333.

[8] 山科亮太，林俊樹，藪田哲郎，“ニューラルネットワークを用いた 1 自由度フレキシブルアームの非線形システム同定と学習制御”，日本機械学会論文集，Vol. 70, No. 693, (2004), pp.1441-1448.

[9] 池野谷康司, 山科亮太, 原正之, 藪田哲郎, “1 自由度フレキシブルアームの反復学習制御”, 計測自動制御学会論文集，Vol. 41, No. 1, (2005), pp.91-93.

[10] 山科亮太，林俊樹，藪田哲郎，“ニューラルネットワークを用いた 1 自由度フレキシブルアームの学習制御”，第 21 回日本ロボット学会学術講演会講演概要集，(2003), 1C16.

遺伝的アルゴリズム（ＧＡ）

[11] GOLDBERG D. E. “Genetic Algorithms in Search”, *Optimization and Machine Learning*, 1989.

[12] 北野宏明，“遺伝的アルゴリズム”，人工知能学会誌，Vol. 7, No. 1, (1992), pp.26-37.

[13] 伊庭斉志，“遺伝的プログラミングと進化論的な学習 (<小特集>「遺伝的アルゴリズムの

新しい潮流」)”，人工知能学会誌，No. 9, Vol. 4, (1994), pp.512-517.
[14] 馬場則夫，久保田直行，“遺伝アルゴリズムを用いたロボットマニピュレータの軌道生成及び障害物回避”，日本ロボット学会誌，Vol. 11, No. 2, (1993), pp.299-302.
[15] 星野力，光本大輔，長野徹，“ロボット行動の進化とその頑健性”，計測自動制御学会論文集，Vol. 33, No. 6, (1997), pp.533-540.
[16] 柴田崇徳, 福田敏男，“Genetic Algorithm を用いた移動ロボットの最適経路計画 ： 第 2 報，複数ロボットのための利己的計画と協調的計画”，日本機械学会論文集. C 編，Vol. 59, No. 560, (1993), pp.1134-1141.
[17] 遠藤謙，川内野明洋，前野隆司，“進化的計算法を用いたリンク型移動ロボットの形態と運動パターンのデザイン法”，日本ロボット学会誌，Vol. 22, No. 2, (2004), pp.273-280.
[18] 小島宏行，“遺伝的アルゴリズムを用いた CP 制御フレキシブルロボットアームの軌道計画(機械力学，計測，自動制御)”，日本機械学会論文集．C 編，Vol. 68, No. 670, (2002), pp.1784-1790.
[19] 小島宏行，木部哲治，“遺伝的アルゴリズムを用いた最適軌道計画による 2 関節フレキシブルロボットアームの残留振動制制御”，日本ロボット学会誌，Vol. 19, No. 7, (2001), pp.905-912.
[20] 柴田崇徳，“マルチエージェントシステムの遺伝アルゴリズムを用いた学習による進化的協調行動”，日本ロボット学会誌，Vol. 11, No. 8, (1993), pp.168-176.
[21] 田川聖治，川口俊介，井上克巳，羽根田博正，“遺伝的アルゴリズムとアフォーダンスを用いた知能ロボットの創発”，日本ロボット学会誌，Vol. 17, No. 7, (1999), pp.1023-1030.
[22] MINAMI M, “Manipulator visual servoing and tracking of fish using a genetic algorithm”, *An international Journal of Industrial Robot*, Vol. 26, No. 4, (1999), pp.278-289.

サポートベクターマシン

[23] CORTES C, “Support Vector Networks”, *Machine Learning*, No. 20, (1995), pp.273-297.
[24] 津田宏治，“サポートベクターマシンとは何か”, 電子情報通信学会誌, Vol. 83, No. 6, (2000), pp.460-466.
[25] TONG S, “Support vector machine active learning with applications to text classification”, *In Journal of Machine Learning Research*, No. 2, (2001), pp.45-66.
[26] BURGES C, “A Tutorial on Support Vector Machines for Pattern Recognition,” Data Mining and Knowledge Discovery, 2(2), 1998, 121-167.
[27] 三浦純，森田英夫，ヒルドミヒャエル，白井良明，“SVM による物体と位置の視覚学習に基づく屋外移動ロボットの位置推定”，Journal of Robotics Society of Japan, Vol. 25, No. 5, (2007), pp.792-798.

ブースティング

[28] フロインドヨアブ，シャピリロバート，安倍直樹，“ブースティング入門 (<特集>計算学習理論の進展と応用可能性)”，人工知能学会誌，Vol. 14, No. 5, (1999), pp.771-780.
[29] 井谷久博，古橋武，“ブースティングアルゴリズムを用いた自律移動ロボットの制御ルール獲得”，計測自動制御学会論文集，Vol. 43, No. 10, (2007), pp.919-925.
[30] SHAPIRE R. E. “A brief introduction to boosting”, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, (1999), pp.1401-1406.

[31] 村田昇，金森敬文，竹之内高志，“7. ブースティングと学習アルゴリズム : 三人寄れば文殊の知恵は本当か?(<小特集>確率を手なずける秘伝の計算技法-古くて新しい確率・統計モデルのパラダイム-)”，電子情報通信学会誌，Vol. 88, No. 9, (2005), pp.724-729.

クラスタリング

[32] JAIN AK, “Data clustering : a review”, *ACM Computing Surveys*, No. 31, (1999), pp.264-323.
[33] 宮永喜一，奥村伸二，栃内香次，“自己組織化クラスタリングの汎化性と適応能力について”， 電子情報通信学会論文誌. A，基礎・境界，Vol. 75, No. 7, (1992), pp.1207-1215.
[34] 呂建軍，時永祥三，“遺伝的プログラミングによる時系列モデルの集合的近似とクラスタリングへの応用(ディジタル信号処理)”，電子情報通信学会論文誌. A，基礎・境界，J88-A, No.7, (2005), pp.803-813.

データマイニング

[35] 神蔦敏弘，“データマイニング分野のクラスタリング手法(1)：クラスタリングを使ってみよう!”，人工知能学会誌，Vol. 18, No. 1, (2003), pp.59-65.
[36] 元田浩，鷲尾隆，“機械学習とデータマイニング (<特集> 大規模データベースからの知識獲得)”，人工知能学会誌，Vol. 12, No. 4, (1997), pp.505-512.
[37] AGRAWAL R, “Database Mining: A Performance Perspective”, *IEEE Trans. Knowledge and Data Engineering*, Vol. 5, No. 6, (1993), pp.914-925.
[38] CHEN M.S, “Data Mining: An Overview from a Database Perspective”, *IEEE Trans. Knowl. Data Eng*, Vol. 8, No. 6, (1996), pp.866-883.

＜強化学習一般論＞

[39] Sutton, R. S. & Barto, A., “Reinforcement Learning: An Introduction”, *A Bradford Book, The MIT Press*, 1998.
[40] 三上貞芳，皆川雅章 訳，“強化学習”，森北出版株式会社，(2000), (Richard S. Sutton and Andrew G. Barto: ”Reinforcement learning: An Introduction”, MIT Press/Bradford Books, March (1998))
[41] 畝見達夫，“強化学習 (<小特集>「最近の機械学習」) Reinforcement Learning (<Special Issue> Recent Machine Learning)”，人工知能学会誌，Vol. 9, No. 6, (1994), pp.830-836.
[42] 畝見達夫，“強化学習法とロボットへの応用 Reinforcement Learning Method and its Applications to Robot”，日本ロボット学会誌，Vol. 13, No. 1, (1995), pp.51-56.

＜強化学習各種理論＞

動的計画法

[43] Bertsekas, D. P., “Dynamic Programming and Optimal Control”, *Athena Scientific, Belmont, MA*, (1995).
[44] Bertsekas, D. P., Tsitsiklis, J. N., “Neuro-Dynamic rogramming”, *Athena Scientific, Belmont, MA*, (1996).
[45] Dreyfus, S. E., Law, A. M., “The Art and Theory of Dynamic Programming”, *Academic Press, New York*, 1977.
[46] Ross, S., “Introduction to Stochastic DynamicProgramming”, *Academic Press, New York*, (1983).

[47] White, D. J., “Dynamic Programming”, *Holde-Day, San Francisco*, (1969).
[48] Whittle, P., “Optimaization over Time”, *Wiley, New York*, vol. 1, (1982).
[49] Whittle, P., “Optimaization over Time”, *Wiley, New York*, vol. 2, (1983).
[50] Kumar, V., Kanal, L. N., “The CDP:A unifying formulation for heuristic search, dynamic programming, and branch-and-bound”, *In L. N. Kanal and V. Kumar (eds.), Search in Artificial Intelligence*, (1988), pp.1-37.
[51] Minsky, M. L., “Steps toward artificial intelligence”, *Proceedings of the Institute of Radio Engineers, 49:8-30. Reprinted in E. A. Feigenbaum and J. Feldman (eds.), Computers and Thoughe*, (1961), pp.406-450.
[52] Watkins, C. J. C. H., “Learning from Delayed Rewards”, *Ph.D. thesis, Cambridge University*, (1989).

モンテカルロ法

[53] Curtiss, J. H., “A theoretical comparison of the efficiencies of two classical methods and a Mote Carlo method for computing one component of the solution of a set of linear algebraic equations”, *In H. A. Meyer (ed.), Symposium on Mote Carlo Methods, Wiley, New York*, (1954), pp.191-233.
[54] Rubinstein, R. Y., “Simulation and the Monte Carlo Method”, *Wiley, New York*, (1981).
[55] Michie, D., and Chambers, R. A., “BOXES: An experiment in adaptive control”, *In E. Dale and D. Michie (eds.), Machine Intelligence 2, Oliver and Boyed, Edinburgh*, (1968), pp.137-152.
[56] Narendra, K. S., and Wheeler, R. M., “Decentralized learning in finite Markov chains”, *IEEE Transactions on Automatic Control*, AC31, No. 6, (1986), pp.519-526.
[57] Barto, A. G., and Duff, M., “Monte Carlo matrix inversion and reinforcement learning”, *In J. D. Cohen, G. Tesauro, and J. Alspector (eds.), Advances in Neural Information Processing Systems: Proceedings of the 1993 Conference, Morgan Kaufmann, San Francisco*, (1987), pp.687-694.
[58] Curtiss, J. H., “A theoretical comparison of the efficiencies of two classical methods and a Mote Carlo method for computing one component of the solution of a set of linear algevraic equations”, *In H. A. Meyer (ed.), Symposium on Mote Carlo Methods, Wiley, New York*, (1954), pp.191-233.
[59] Singh, S. P., and Sutton, R. S., “Reinforcement learning with replacing eligibility traces”, *Machine Learning*, No.22, (1996), pp.123-158.

TD学習

[60] Samuel, A. L., “Some studies in machine learning using the game of checkers”, *IBM Journal on Research and Development, 3:211-229 Reprinted in E. A. Feigenbaum and J. Feldman (eds.), Computers and Thought, McGraw-Hill, New York*, (1963), pp.71-105.
[61] Klopf, A. H., “Brain function and adaptive systems -A heterostatic theory”, *Technical Report AFCRL-72-0164, Air Force Cambridge Research Laboratories, Bedford, MA. A summary appears in Proceedings of the International Conference on Systems, Ma, and Cybernetics, IEEE Systems, Man, and Cybernetics Society, Dallas, TX*, (1974).
[62] Holland, J. H., “Adaptation in Natural and Artificial Systems”, *University of Michigan Press, Ann Arbor*, (1975).
[63] Booler, L. B., “Intelligent Behavior as an Adaptation to the Task Environment”, *Ph.D. thesis, University of Michigan, Ann Arbor*, (1982).
[64] Sutton, R. S., “Learning to predict by the method of temporal differences”, *Machine Learning*, No. 3, (1988), pp.9-44.

Sarsa

[65] Rummery, G. A., and Niranjan, M., "On-line Q-learning using connectionist systems", *Technical Report CUED/F-INFENG/TR 166, Engineering Department, Cambridge University*, (1994).

[66] Holland, J. H., "Escaping brittleness: The possibility of general-purpose learning algorithms applied to rule-based system", *In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (eds.), Machine Learning: An Artificial Intelligence Approach, vol. 2, Morgan Kaufmann, San Mateo, CA*, (1986), pp.593-623

[67] Wilson, S. W., "ZCS: A zeroth order classifier system", *Evolutionary Computation*, No. 2, (1994), pp.1-18.

Q-Learning

[68] Watkins, C. F. C. H., "Learning from Delayed Rewards", *Ph.D. thesis, Cambridge University*, (1989).

[69] Watkins, C. F. C. H., and Dayan, P., "Q-learning", *Machine Learning*, No. 8, (1992), pp.279-292.

[70] Dayan, P., "The convergence of TD(λ) for general λ", *Machine Learning*, No. 8, (1992), pp.341-362.

[71] Jaakkola, T., Jordan, M. I., and Singh, S. P., "On the convergence of stochastic iterative dynamic programming algorithms", *Neural Computation*, No. 6, (1994), pp.1185-1201.

[72] Tsitsiklis, J. N., "Asynchronous stochastic approximation and Q-learning", *Machine Learning*, No. 16, (1994), pp.185-202.

Actor-Critic

[73] Witten, I. H., "An adaptive optimal controller for discrete-time Markov environments", *Information and Control*, No. 34, (1977), pp.286-295.

[74] Barto, A. G., Sutton, R. S., and Anderson. C. W., "Neuronlike elements that can solve difficult learning control problems", *IEEE Transactions on Systems, Man, and Cybernetics, 13:835-846. Reprinted in J. A.Anderson and E. Rosenfeld (eds.), Neurocomputing: Foundations of Research*, (1988), pp. 535-549.

[75] Barto, A. G., "Adaptive critics and the basal ganglia", *IN J. C. Houk, J. L. Davis, and D. G. Beiser (eds.), Models of Information Processing in the Basal Ganglia, MIT Press, Cambridge, MA*, (1995), pp.215-232.

[76] Houk, J. C., Adams, J. L., and Barto, A. G., "A model of how the basal ganglia generates and uses neural signals that predict reinforcement", *In J. C. Houk, J. L. Davis, and D. G. Beiser (eds.), Models of Information Processing in the Basal Ganglia, MIT Press, Cambridge, MA*, (1995), pp.249-270.

R学習

[77] Schwartz, A., "A reinforcement learning method for maximizing undiscounted rewards", *In Proceedings of the Tenth International Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA*, (1993), pp.298-305.

[78] Mahadevan, S., "Average reward reinforcement learning: Foundations, algorithms, and empirical results", *Machine Learning*, No. 22, (1996), pp. 159-196.

[79] Tadepalli, P., and Ok, D., "H-learning: A reinforcement learning method to optimize undiscounted average reward", *Technical Report 94-30-01. Oregon State University. Computer Science Department, Corvallis*, (1994).

DP,MC,TDの関連

[80] Werbos, P. J., “Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research”, *IEEE Transactions on Systems, Man, and Cybernetics*, No. 17, (1987), pp.7-20.

Profit- Sharing

[81] 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(<特集>計算学習理論の進展と応用可能性)”，人工知能学会誌，Vol. 14, No. 5, (1999), pp.800-807.

[82] 荒井幸代，宮崎和光，小林重信，“マルチエージェント強化学習の方法論 :Q-learning と Profit Sharing による接近”，人工知能学会誌，Vol. 13, No. 5, (1998), pp.609-618.

[83] 堀内匡，藤野昭典，片井修，椹木哲夫，“経験強化を考慮した Q-Learning の提案とその応用”，計測自動制御学会論文集，Vol. 35, No. 5, (1999), pp.645–653.

＜強化学習の連続系・動的問題への応用＞

[84] 森本淳，銅谷賢治，“強化学習を用いた高次元連続状態空間における系列運動学習 : 起き上がり運動の獲得”，電子情報通信学会論文誌. D-II，情報・システム，II-パターン処理，J82-D-II, No. 11, (1999), pp.2118-2131.

[85] 釜谷博行，阿部健一，“連続状態空間のための強化学習アルゴリズム”，八戸工業高等専門学校紀要，No. 42, (2007), pp.65-68.

[86] 佐藤仁樹，“高次元連続状態空間における強化学習-多変量解析による状態空間の圧縮”，電子情報通信学会技術研究報告，Vol. 105, No. 547, (2006), pp.7-12.

[87] 柴田聡志，神谷昭基，“強化学習の連続値への適用”，釧路工業高等専門学校紀要，No. 41, (2007), pp.39-45.

[88] 川田誠一, 谷村暁之, 小口俊樹,“連続環境における移動ロボットの強化学習”, JSME annual meeting, No. 5, (2003), pp.239-240.

[89] 櫻井義尚，本多中二，“連続な状態行動空間において近傍状態の報酬予測を用いた強化学習（エージェント・学習）”，情報処理学会研究報告. MPS，数理モデル化と問題解決研究報告，No. 130, (2004), pp.61-64.

[90] 有江浩明，尾形哲也，谷淳，菅野重樹，“CTRNN を用いた連続な状態空間における強化学習法の提案”，ロボティクス・メカトロニクス講演会講演概要集，No. 1, (2006). 1K32.

[91] 田中昭雄，中田洋平，松本隆，“動的環境下の強化学習アルゴリズム : Sequential Monte Carlo とサンプル初期化”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 104, No. 759, (2005), pp.101-106.

[92] 野田五十樹,“動的環境における強化学習のステップサイズパラメータ調整法（強化学習）”，情報処理学会研究報告. ICS，[知能と複雑系]，No. 104, (2008), pp.73-80.

[93] 高橋哲也，安達雅春，“強化学習における環境変化の検出法”，電子情報通信学会技術研究報告. NLP，非線形問題，Vol. 104, No. 583, (2005), pp.35-40.

＜強化学習と他学習の組み合わせ＞

[94] SAMEJIMA K, OMORI T, “Adaptive internal state space construction method for reinforcement learning of a real-world agent”, *Neural networks: the official journal of the International Neural Network Society*, Vol. 12, No. 7, (1999), pp.1143-1155.

[95] 柴田克成，岡部洋一，伊藤宏司，“ニューラルネットワークを用いた Direct-Vision-Based 強化学習-センサからモータまで-”，計測自動制御学会論文集，Vol. 37, No. 2, (2001), pp.168-177.

[96] 神尾武司，曽我咲十美，三堀邦彦，“ファジィ ART ニューラルネットワークによる強化学習のための状態空間の構成法”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 103, No. 734, (2004), pp.43-48.

[97] 森田昌彦，新保智之，蓮尾高志，山根健，“選択的不感化ニューラルネットを用いた強化学習の効率化”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 107, No. 542, (2008), pp.355-359.

[98] 中村太亮，神尾武司，三堀邦彦，藤坂尚登，“ART ニューラルネットワークによる強化学習のための状態生成器の改良”, 電子情報通信学会技術研究報告. NLP, 非線形問題, Vol. 105, No. 125, (2005), pp.25-30.

[99] 井上勇気，赤塚洋介，佐藤裕二，“強化学習における報酬値探索への GA の適用”，The 21st Annual Conference of the Japanese Society for Artificial Intelligence，(2007), 3D9-1.

[100] 矢野 友貴, 柴田 剛志, 横山 大作, 田浦 健次朗, 近山 隆, “GA と TD(λ)学習の組み合わせによるゲーム局面評価パラメータの調整(学習 1),” 情報処理学会研究報告. GI, [ゲーム情報学], 2009(27), 63-70.

[101] 伊藤一之，松野文俊，“GA により探索空間の動的生成を行う Q 学習による実多自由度ロボットの制御 : 階層構造の拡張と蛇型ロボットへの適用”，日本ロボット学会誌，Vol. 21, No. 5, (2003), pp.526-534.

[102] 亀井圭史，石川眞澄，“遺伝的アルゴリズムによる移動ロボットの強化学習パラメータ最適化”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 104, No. 759, (2005), pp.119-124.

[103] 土谷千加夫，木村元，佐久間淳，小林重信，“重点サンプリングを用いた GA による強化学習”，人工知能学会論文誌，Transactions of the Japanese Society for Artificial Intelligence : AI, No. 20, (2005), pp.1-10.

[104] 米井友浩, 村川正宏, 吉澤修治, “遺伝的アルゴリズムを用いた時変環境における Q-learning”, 電子情報通信学会技術研究報告. NC, ニューロコンピューティング, Vol. 97, No. 448, (1997), pp.71-78.

[105] 櫻井義尚，鶴田節夫，“強化学習を用いた進化的アルゴリズムのパラメータ学習”，情報処理学会研究報告. MPS，数理モデル化と問題解決研究報告，MPS-75(5), (2009), pp.1-6.

[106] 伊藤一之，松野文俊，“GA により探索空間の動的生成を行う Q 学習”，人工知能学会誌，Vol. 16, No. 6, (2001), pp.510-520.

[107] LIKAS A, “A reinforcement learning approach to on-line clustering”, *Neural Comput*. Vol. 11, No. 8, (1999), pp.1915-1932.

[108] 北越大輔，山口晃昌，塩谷浩之，中野良平，“クラスタリングを用いた強化学習システム IPMBN の環境変化への適応について(ニューラルネットワーク画像復元及び一般)”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 106, No. 500, (2007), pp.65-70.

[109] 小谷直樹，布引雅之，谷口研二，“強化学習における状態数を抑制するクラスタリング方法”，システム制御情報学会論文誌，Vol. 22, No. 1, (2009), pp.21-28.

＜強化学習 アプリケーション(1)上位層レベル問題＞

[110] 吉岡琢，石井信，“EM アルゴリズムによるオセロの評価関数の学習”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 98, No. 219, (1998), pp.85-92.
[111] 星野孝総，亀井且有，“ファジィ環境評価ルールを用いた強化学習の提案とチェスへの応用”，日本ファジィ学会誌，Vol. 13, No. 6, (2001), pp.626-632.
[112] 金田道明，長尾智晴，“強化学習を用いた将棋における事例の獲得”，電子情報通信学会総合大会講演論文集. 情報・システム，No. 1, (1999), pp.169.
[113] 佐々木宣介，“機械学習と自動プレイを用いた将棋類の類似度比較について”，情報処理学会研究報告. GI，[ゲーム情報学]，No. 23, (2006), pp.41-48.
[114] 伊藤昭, 大橋資紀, 寺田和憲,“非零和ゲームの強化学習 : 相手の行動を読むプログラム”, 知識ベースシステム研究会，No. 71, (2005), pp.53-60.
[115] 山崎善正，石川眞澄，“強化学習を用いた移動ロボットの行動制御”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 101, No. 735, (2002), pp.83-90.
[116] 竹口知男，小尻一憲，大橋美奈子，今井弘之，能勢和夫，“強化学習によるロボット移動経路の探索”，関西支部講演会講演論文集，No. 76, (2001), "5-47"-"5-48".
[117] 甲斐孝史，石川眞澄，“強化学習を用いた変動環境下の最短経路探索”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 109, No. 461, (2010), pp.119-124.
[118] 奥本泰久，小川隆寿，“強化学習法を用いた自動溶接機の移動経路最適化”，日本船舶海洋工学会論文集，No. 5, (2007), pp.85-89.
[119] 三堀邦彦，神尾武司，田中隆博，“潮流の影響を考慮した操船運動の強化学習について”，電子情報通信学会技術研究報告. NLP，非線形問題，Vol. 102, No. 142, (2002), pp.27-32.
[120] 神尾武司，森賢児，藤坂尚登，三堀邦彦，“適応的な状態空間分割を考慮した強化学習による操船経路決定”，電子情報通信学会技術研究報告. NLP，非線形問題，Vol. 104, No. 754, (2005), pp.25-30.
[121] 安信誠二，“fPSP-強化学習による駐車支援知識の獲得”，インテリジェントシステム・シンポジウム講演論文集，No. 12, (2002), pp.105-108.
[122] 五十嵐陽，三堀邦彦，“強化学習アルゴリズムに基づく 4 輪車の経路決定と切り返しの導入”，電子情報通信学会技術研究報告. NLP，非線形問題，Vol. 108, No. 442, (2009), pp.1-6.
[123] 小池康晴，銅谷賢治，“強化学習による自動車運転技能の獲得”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 96, No. 584, (1997), pp.107-114.
[124] 小池康晴, 銅谷賢治,“マルチステップ状態予測を用いた強化学習によるドライバモデル”, 電子情報通信学会論文誌．D-II，情報・システム，II-パターン処理，J84-D-II(2)，(2001), pp.370-379.
[125] 参沢匡将，木村春彦，広瀬貞樹，大里延康，“強化学習型マルチエージェントによる交通信号制御”, 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, J83-D-I, No. 5, (2000), pp.478-486.
[126] 蒋励，藤田聡，“小型乗り合いバスシステムにおける最適発車間隔問題のモデル化とその

強化学習による獲得手法の提案”，情報処理学会研究報告. MPS，数理モデル化と問題解決研究報告，No. 20, (2003), pp.35-38.
[127] 小越康宏, 木村春彦, 広瀬貞樹, 大里延康,“エレベータ群管理システムに対する一考察”, 電子情報通信学会論文誌. A，基礎・境界，J84-A, No. 1, (2001), pp.22-32.
[128] 松井藤五郎，大和田勇人，“強化学習を用いた株式取引シミュレーション”，情報科学技術フォーラム一般講演論文集，Vol. 5, No. 2, (2006), pp.257-258.
[129] 浅田稔，野田彰一，俵積田健，細田耕，“視覚に基づく強化学習によるロボットの行動獲得”，日本ロボット学会誌，Vol. 13, No. 1, (1995), pp.68-74.

＜強化学習 アプリケーション(2)下位層レベル問題＞

[130] K. Doya, “Reinforcement learning in animals and robots”, *Proc. International Workshop on Brainware*, (1996), pp.69–71.
[131] H. Kimura and K. Kobayashi, “Reinforcement learning using stochastic gradient algorithm and its application”, *IEE Japan Trans. on Electronics, Information and Systems*, Vol. 119-C, No. 8, (1999), pp.931–934.
[132] K. Ito, T. Kamegawa, and F. Matsuno, “Extended QDSEGA for controlling real robots–acquisition of locomotion patterns for snake-like robot”, *Proceedings of the 2003 IEEE International Conference on Robotics and Automation*, vol.1, (2003), pp.791–796.
[133] 小林祐一，加藤真人，細江繁幸，“分解能の調整可能な画像情報からの状態空間の構成”，日本ロボット学会誌，Vol. 25, No. 5, (2007), pp.770-778.
[134] 浅水喬大，小林祐一，“運動学習と動作計画に基づいたロボットの身体と対象物表現の獲得”，日本ロボット学会学術講演会予稿集，27th, (2009), 1L1-04.
[135] 野村貴洋，浅水喬大，小林祐一，“自己組織化マップと強化学習を用いた冗長マニピュレータの動作獲得”，知能システムシンポジウム資料，36th, (2009), pp.107-110.
[136] 小林祐一，藤井博基，細江繁幸，“一次元空間への低次元化写像を用いた対象物操作の強化学習”，計測自動制御学会論文集，Vol. 42, No. 7, (2006), pp.814-821.
[137] 小林祐一，細江繁幸，“2 次元空間への低次元化写像を用いた対象物操作強化学習のための関数近似”，知能システムシンポジウム資料，No. 32, (2005), pp.421-424.
[138] 河原井伸行，小林祐一，“行動空間に拘束条件のある対象物操作の制御則獲得”，計測自動制御学会システム・情報部門学術講演会講演論文集，(2009), 2C4-4.
[139] 高崎雄太，河原井伸行，小林祐一，“SVM と補間を利用した対象物操作学習のための接触モード境界推定”，日本機械学会ロボティクス・メカトロニクス講演会講演論文集，(2009), 2A2-E05.
[140] Y.Kobayashi, M.Shibata, S.Hosoe, Y.Uno, “Learning of Object Manipulation with Stick/Slip Mode Switching”, *Proc. of Int. Conf. on Intelligent Robots and Systems*, (2008), pp.373-379.
[141] 小林祐一, 湯浅秀男, 細江繁幸,“強化学習のための矩形基底による自律分散型関数近似”, 計測自動制御学会論文集，Vol. 40, No. 8, (2004), pp.849-858.
[142] Y. Kobayashi, H Fujii, S. Hosoe, “Reinforcement Learning for Manipulation Using Constraint between Object and Robot”, *IEEE Int. Conf. on Systems, Man & Cybernetics Hawaii USA*, (2005), pp.871-876.
[143] 鴇田正俊, 福田敏男, 中根幹子,“強化学習によるマニピュレータの最適接近速度の学習”,

木更津工業高等専門学校紀要，No. 33, (2000), pp.7-14.
[144] 西村政哉，吉本潤一郎，時田陽一，中村泰，石井信，“複数制御器の切換学習法による実アクロボットの制御”, 電子情報通信学会論文誌, 基礎・境界, J88-A, No. 5, (2005), pp.646-657.
[145] 泉田啓，真野修輔，“強化学習における状態空間の縮小法について”，ロボティクス・メカトロニクス講演会講演概要集，(2003), 2A1-3F-B3.
[146] 井沢淳，近藤敏之，伊藤宏司，“階層構造を利用した強化学習によるダイナミックマニピュレーション”，SICE Symposium on Decentralized Autonomous Systems, No. 14, (2002), pp.291-296.
[147] 柴田克成，杉坂政典，伊藤宏司，“強化学習によるリーチング動作の獲得”，電子情報通信学会技術研究報告．NC，ニューロコンピューティング，Vol. 100, No. 688, (2001), pp.107-114.
[148] 児島史周，片田喜章，Svinin Mikhail，上田完次，“学習を用いた複数アーム型ロボットによる物体の協調持ち上げ作業”，ロボティクス・メカトロニクス講演会講演概要集，(2000), 2P1-33-037.
[149] 山田訓，“RBF 型リカレントネットを用いた強化学習”，電子情報通信学会総合大会講演論文集，情報・システム，No. 1, (1996), pp.25.
[150] 山田訓，渡邊彰，塩野悟，“強化学習によるマニピュレータの制御”，電子情報通信学会総合大会講演論文集，情報・システム，No. 1, (1995), pp.69.
[151] 成瀬継太郎，嘉数侑昇，“強化学習を用いた分散学習エージェントによるマニピュレータのリアクティブプランニング”，日本機械学会論文集 総務偏，Vol. 61, No. 581, (1995), pp.131-137
[152] 吉本潤一郎，石井信，佐藤雅昭，“オンライン EM アルゴリズムによる強化学習法の acrobot 制御への応用”，電子情報通信学会論文誌. D-II，情報・システム，II-パターン処理，J83-D-II, No. 3, (2000), pp.1024-1033.
[153] 前田雄介，坂本記章，“マルチエージェント型組立ロボットシステムのための強化学習に基づく作業割当”，日本機械学会創立 110 周年記念 2007 年度年次大会講演論文集，Vol. 7, (2007), pp.263-264.
[154] 岡本太一，小林祐一，大西正輝，“ロボットの動作生成のための画像特徴の獲得”，日本ロボット学会学術講演会，27th, (2009), 1L1-05.
[155] 河原井伸行，小林祐一，“マニピュレータを用いた対象物の抱きかかえ操作における制御の獲得”，日本ロボット学会学術講演会予稿集，26th, (2008), 3K1-01.
[156] 高崎雄太，河原井伸行，小林祐一，“マニピュレータを用いた対象物操作学習のための接触モード境界推定”，知能システムシンポジウム資料，36th, (2009), pp.337-342.
[157] 栗田英介，里悠太，小林祐一，“拡散学習を用いた不完全な知覚を有するロボットのための状態推定法”，日本機械学会ロボティクス・メカトロニクス講演会講演論文集，(2009), 1A1-F16.
[158] 岡本太一，浅水喬大，小林祐一，大西正輝，“距離画像からの身体の抽出を用いたマニピュレータによるリーチング行動の学習”，日本ロボット学会学術講演会予稿集，26th, (2008), 3F1-03.
[159] 小林祐一，相山康道，井上康介，ＺＨＵ Ｃ，新井民夫，“マニピュレータによる押し操作・弾き操作の獲得”，日本ロボット学会学術講演会予稿集，15th，特殊号:第 1 分冊，(1997), pp.159-160.

[160] 成瀬継太郎，嘉数侑昇，“学習オートマトンによる冗長マニピュレータのパスプランニングの戦略獲得”，情報処理学会第 45 回全国大会，No. 3, (1992), pp.61-62.

＜強化学習 アプリケーション(3)生物型ロボットへの適用＞

[161] 牧野研司，中村泰，柴田智広，石井信，“CPG-Actor-Critic 法によるミミズ型ロボットの推進運動の獲得”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 106, No. 588, (2007), pp.203-208.
[162] 世戸大地，小川賢一，嵯峨宣彦，佐藤俊之，高梨宏之，長南征二，“蠕動運動型ロボットにおける強化学習による運動パターンの獲得”，ロボティクス・メカトロニクス講演会講演概要集，(2006), 2P1-A22.
[163] 伊藤一之，高山明宏，“身体と環境の特性を利用した状態-行動空間の抽象化：強化学習を用いた自律ヘビ型ロボットへの適用”，日本知能情報ファジイ学会誌，Vol. 21, No. 3, (2009), pp.402-410.
[164] 中西恒平，大砂古悠生，中西智士，堤一義，“腰部関節を有するヤモリ型ロボットの強化学習に基づく歩行獲得”, ロボティクス･メカトロニクス講演会講演概要集, (2010), 2P1-F24.
[165] 陶衛軍，王碩玉，河田耕一，四宮葉一，石田健司，木村哲彦，“強化学習による 4 足ロボットの安定歩行獲得”，日本機械学会中国四国支部総会・講演会講演論文集，No. 44, (2006), pp.439-440.
[166] 山口明彦，高松淳，小笠原司，“強化学習によるロボットの動作獲得のための基底関数に基づく行動空間生成手法 DCOB：実機多自由度ロボットの匍匐動作への適用”，ロボティクス・メカトロニクス講演会講演概要集，(2010), 2P1-G10.
[167] 綿貫啓一，新村弘樹，“強化学習を用いた昆虫規範型多足ロボットの歩容の最適化”，日本機械学会関東支部総会講演会講演論文集，No. 9, (2003), pp.127-128.
[168] 三上貞芳，田野浩明，嘉数侑昇，“強化学習による多足歩行ロボットの適応的歩様獲得に関する研究”，日本機械学会論文集. C 編，Vol. 60, No. 580, (1994), pp.4252-4259.
[169] 守田観輝夫，石川眞澄，“強化学習を用いた生存欲に基づく行動の創発”，電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 108, No. 480, (2009), pp.279-283.
[170] 清水一貴，山田訓，“ゴム人工筋制御の強化学習,”電子情報通信学会技術研究報告．NC，ニューロコンピューティング，Vol. 108, No. 480, (2009), pp.297-300.
[171] 木村元, 小林重信,“ロボットアームのほふく行動の強化学習 ： 確率的傾斜法による接近”, 人工知能学会誌，Vol. 14, No. 1, (1999), pp.122-130.
[172] 丸山淳一，松原崇充，HALE Joshua G，森本淳，“強化学習を用いたヒューマノイドロボットによる転倒回避ステップ動作の学習”，日本ロボット学会誌，Vol. 27, No. 5, (2009), pp.527-537.
[173] 川島諒，元木誠，小圷成一，平田廣則，“強化学習を用いた二足歩行ロボットの行動選択の最適化”，電気学会電子・情報・システム部門大会講演論文集，(2008), OS7-18.
[174] 長沼輝樹，高村松三，“強化学習とニューラルネットワークによる 2 足歩行ロボットの歩行軌道の学習”, 日本機械学会北陸信越支部総会講演会講演論文集, 42nd, (2005), pp.219-220.
[175] 吉田利光，高村松三，“強化学習と自己組織化マップを用いた 2 足歩行ロボットの歩行制御”，日本機械学会北陸信越支部総会講演会講演論文集，42nd, (2005), pp.217-218.

[176] 伊藤一之，松野文俊，“GA により探索空間の動的生成を行う Q 学習による実多自由度ロボットの制御 : 階層構造の拡張と蛇型ロボットへの適用”，日本ロボット学会誌，Vol. 21, No. 5, (2003), pp.526-534.
[177] K. Ito, F. Matsuno, “Control of hyper-redundant robot using QDSEGA”, *SICE 2002. Proceedings of the 41st SICE Annual Conference 3*, (2002), pp.1499-1504.
[178] K. Ito, T, Kamegawa, F, Matsuno, “Extended QDSEGA for Controlling Real Robot : Acquisition of Locomotion Patterns for Snake : like Robot”, *Robotics and Automation 1*, (2003), pp.791-796.
[179] 石黒章夫，市川真吾，久保敷悟，武藤勝彦，内川嘉樹，“免疫ネットワークを用いた 6 脚歩行ロボットの自律分散的歩容制御 : 強化学習による歩容生成の一手法”，日本機械学会論文集. C 編，Vol. 63, No. 609, (1997), pp.1679-1684.
[180] 森健，中村泰，石井信，“二足歩行運動に対する方策勾配法に基づいた強化学習法”電子情報通信学会技術研究報告. NC，ニューロコンピューティング，Vol. 103, No. 734, (2004), pp.73-78.
[181] 本山晴寿，山科亮太，原正之，黄健，藪田哲郎，“強化学習によって獲得される芋虫型ロボットの前進行動形態に関する考察”，日本機械学会論文集. C 編，Vol. 72, No. 723, (2006), pp.3525-3532.
[182] 本山晴寿，山科亮太，黄健，藪田哲郎，“強化学習を用いたロボットの前進行動形態に関する考察”，第 11 回ロボティクスシンポジア予稿集，(2006), pp.258-263.
[183] 鄭英美，井上将志，原正之，黄健，藪田哲郎，“強化学習による二次元移動ロボットの行動獲得とその学習知識の操作”，日本機械学会論文集. C 編，Vol. 75, No. 749, (2009), pp.122-131.
[184] 坂井直樹，豊田希，藪田哲郎，“強化学習を用いた 4 足歩行ロボットの行動獲得と解析”，ロボティクス・メカトロニクス講演会講演概要集，(2010), 2P1-F23.
[185] 坂井直樹，原正之，藪田哲郎，“強化学習を用いたロボットの大車輪運動の獲得に関する研究”，ロボティクス・メカトロニクス講演会講演概要集，(2009), 2A2-D17.
[186] 川辺直人，原正之，黄健，藪田哲郎，“強化学習によるスポーツロボットの大車輪運動に関する研究”，日本機械学会年次大会講演論文集，5th, (2008), pp.167-168.

＜強化学習 報酬の与え方（客観報酬）＞

[187] 宮崎和光，山村雅幸，小林重信，“強化学習における報酬割当ての理論的考察”，人工知能学会誌，Vol. 9, No. 4, (1994), pp.580-587.
[188] 宮崎和光，山村雅幸，小林重信，“MarcoPolo : 報酬獲得と環境同定のトレードオフを考慮した強化学習システム”，人工知能学会誌，Vol. 12, No. 1, (1997), pp.78-89.
[189] 森山甲一, 沼尾正行, “環境状況に応じて自己の報酬を操作する学習エージェントの構築”, 人工知能学会論文誌 = Transactions of the Japanese Society for Artificial Intelligence, AI 17, (2002), pp.676-683.
[190] 荒井幸代，田中信行，“マルチエージェント連続タスクにおける報酬設計の実験的考察 : RoboCup Soccer Keepaway タスクを例として”，人工知能学会論文誌 = Transactions of the Japanese Society for Artificial Intelligence , AI 21, (2006), pp.537-546.
[191] MOORE A. W, “Prioritized sweeping: Reinforcement learning with less data and less time”, *Machine Learning*, No. 13, (1994), pp.103-129.

[192] SINGH S. P, “Reinforcement Learning with Replacing Eligibility Traces”, *Machine Learning*, No. 22, (1996), pp.123-158.
[193] 宮崎和光，荒井幸代，小林重信，“Profit Sharing を用いたマルチエージェントと強化学習における報酬配分の理論的考察”，人工知能学会誌，Vol. 14, No. 6, (1999), pp.1156-1164.
[194] MAHADEVAN S, “Average Reward Reinforcement Learning : Foundations, Algorithms, and Empirical Results”, *Machine Learning*, No. 22, (1996), pp.159-195.
[195] 保知良暢，新谷虎松，伊藤孝行，大囿忠親，“外部評価機構を導入したマルチエージェント強化学習における過去の事象に基づく報酬配分(人工知能，認知科学)”，電子情報通信学会論文誌. D-I，情報・システム，I-情報処理，J87-D-I, No. 12, (2004), pp.1119-1127.
[196] 矢島英明，大倉和博，上田完次，“進化的手法によるエージェント群の強化学習フレーム獲得 : 大域的報酬と個別的報酬のバランスに関する一考察”，インテリジェント・システム・シンポジウム講演論文集 = FAN Symposium : fuzzy, artificial intelligence, neural networks and computational intelligence 10, (2000), pp.291-294.
[197] 江口徹， 関合孝朗，山田昭彦，清水悟，深井雅之，“報酬自動調整機能を備えた強化学習法によるプラント制御技術”，電気学会論文誌. C，電子・情報・システム部門誌 = The transactions of the Institute of Electrical Engineers of Japan. C, A publication of Electronics, Information and System Society, Vol. 129, No. 7, (2009), pp.1253-1263.
[198] 内部英治，銅谷賢治，“複数報酬のもとでの階層強化学習”，日本ロボット学会誌，Vol. 22, No. 1, (2004), pp.120-129.
[199] 山科亮太，本山晴寿，浦川真理子，黄健，藪田哲郎，“報酬変化を用いた強化学習によるロボットの前進行動獲得”, 日本機械学会論文集 C 編, Vol. 72, No. 717, (2006), pp.1574-1581.
[200] 山科亮太，前原晋策，黄健，藪田哲郎，“報酬変化を用いた Q-Learning による実ロボットの前進行動獲得”，第 10 回ロボティクスシンポジア予稿集，4D3, (2005), pp.411-417.
[201] 山科亮太，前原晋策，石川智弘，藪田哲郎，“報酬変化に基づく Q-Learning の収束性に関する検討”，ロボティクス・メカトロニクス講演会'04，(2004), 2A1-L1-26.
[202] 山科亮太，井上将志，浦川真理子，黄健，藪田哲郎，“報酬変化を用いた Q-Learning によるロボットの行動獲得”，ロボット・メカトロニクス講演会'05，(2005), 1A1-S-058.

＜強化学習 報酬の与え方（主観報酬）＞

[203] Thomaz, A.L., Hoffman, G., Breazeal, C, “Reinforcement Learning with Human Teachers: Understanding How People Want to Teach Robots”, *Robot and Human Interactive Communication, ROMAN 2006. The 15th IEEE International Symposium*, (2006), pp352-357.
[204] A. L. Thomaz, C. Breazeal, “Reinforcement Learning with Human Teachers: Evidence of feedback and guidance with implications for learning performance”, *In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, (2006), pp.1000-1005.
[205] 廣川暢一，鈴木健嗣，“コーチングによる報酬関数の動的生成に基づくエージェントの行動学習”，HAI シンポジウム，(2009), 2D-5.
[206] Hirokawa, M., Suzuki, K., “Coaching to Enhance the Online Behavior Learning of a Robotic Agent”, *Lecture Notes in Computer Science*, 6276, (2010), pp.148-157.
[207] 廣川暢一，鈴木健嗣，“コーチングに基づくロボットのオンライン行動学習”，ロボティクス・メカトロニクス講演会，(2009), 2A2-C21.
[208] Ryota Yamashina, Masafumi Kuroda, Tetsuro Yabuta, “Caterpillar Robot Locomotion Based on

Q-Learning using Objective/Subjective Reward", *Proc. of IEEE/SICE International Symposium on System Integration (SII 2011)*, (2011), pp.1311-1316.
[209] 黒田将史, 山科亮太, 藪田哲郎, "主観報酬学習を用いたイモムシ型ロボットの行動獲得", ロボティクス・メカトロニクス講演会'12, (2012), 1A1-D06.

＜主観的評価によるロボット学習（非強化学習）＞

[210] 廣川暢一, 鈴木健嗣, "透過型デバイスを用いたコーチングによるロボットの学習支援", ロボティクス・メカトロニクス講演会講演概要集, (2010), 2A2-F25.
[211] 廣川暢一, 鈴木健嗣, "教示者による学習支援に基づくエージェントのオンライン行動獲得", 人工知能学会論文誌, Vol. 25, No. 6, (2010), pp.694-702.
[212] Momoko Nakatani, Kenji Suzuki and Shuji Hashimoto, "Subjective-Evaluation Oriented Teaching Scheme for a Biped Humanoid Robot", *Proc. of the 2003 IEEE-RAS International Conference on Humanoid Robots (Humanoids2003), CD-ROM Proceedings*, (2003),
[213] TAKAGI H, "interactive Evolutionary computation fusion of the capabilities of EC optimization and human evaluation", *Proceedings of the IEEE*, Vol. 89, No. 9, (2001), pp.1275-1296.
[214] 高木英行, 畝見達夫, 寺野隆雄, "対話型進化計算法の研究動向(<論文特集>対話型進化計算法)", 人工知能学会誌, Vol. 13, No. 5, (1998), pp.692-703.
[215] 大崎美穂, 高木英行, "対話型 EC 操作者の負担低減：評価値予測による提示インタフェースの改善(<論文特集>対話型進化計算法)", 人工知能学会誌, Vol. 13, No. 5, (1998), pp.712-719.
[216] 徳井直生, 伊庭斉志, "対話型進化的計算によるリズムの生成", 人工知能学会全国大会論文集 = Proceedings of the Annual Conference of JSAI 14, (2000), pp.81-82.
[217] Riley, M. "Coaching: An Approach to Efficiently and Intuitively Create Humanoid Robot Behaviors", *Humanoid Robots, 2006 6th IEEE-RAS International Conference*, (2006), pp.567-574.

＜機械と人間の協調制御＞

[218] 大塚弘文, 柴里弘毅, 川路茂保, "コラボレータによる人間-機械系の協調制御", 日本機械学会論文集. C 編, Vol. 73, No. 733, (2007), pp.2576-2582.
[219] 山中絵里, 村上俊之, 大西公平, "等価質量と仮想インピーダンス設定による人間と移動マニピュレータの協調", 電気学会論文誌. D, 産業応用部門誌 = The transactions of the Institute of Electrical Engineers of Japan. D, A publication of Industry Applications Society Vol.123, No. 10, (2003), pp.1227-1233

＜その他参考文献＞

[220] 足立修一, "システム同定の基礎", 東京電機大学出版局, (2009).
[221] 足立修一, "MATLAB による制御のための上級システム同定", 東京電機大学出版局, (2004).
[222] 山科亮太, "強化学習を用いたロボットの行動獲得の研究", 修士論文, (2006).
[223] 山科亮太, "1 自由度フレキシブルアームのシステム同定と学習制御の研究", 学位論文, (2004).

本論文に関連した著者の発表論文

１．学術雑誌掲載論文

[1] **山科亮太**，林俊樹，藪田哲郎，“ニューラルネットワークを用いた 1 自由度フレキシブルアームの非線形システム同定と学習制御”, 日本機械学会論文集C編, Vol. 70, No. 693, (2004), pp.1441-1448.

[2] **山科亮太**，本山晴寿，浦川真理子，黄健，藪田哲郎，“報酬変化を用いた強化学習によるロボットの前進行動獲得”, 日本機械学会論文集 C 編, Vol. 72, No. 717, (2006), pp.1574-1581.

[3] 本山晴寿，**山科亮太**，原正之，黄健，藪田哲郎，“強化学習によって獲得される芋虫型ロボットの前進行動形態に関する考察”日本機械学会論文集 C 編，Vol. 72, No. 723, (2006), pp.3525-3532.

[4] **山科亮太**，黒田将史，藪田哲郎，“主観報酬を用いた強化学習によるイモムシ型ロボットの行動獲得”，日本機械学会論文集 C 編，Vol. 79, No. 798, (2013), pp.366-371.

[5] 黒田将史，**山科亮太**，藪田哲郎，“主観報酬を用いた強化学習における人間の教示特性に関する考察”，日本機械学会論文集 C 編，Vol.79, No.801, (2013), pp.1770-1774.

２．国際会議

[6] Ryota Yamashina, Masafumi Kuroda, Tetsuro Yabuta, Caterpillar Robot Locomotion Based on Q-Learning using Objective/Subjective Reward, *Proc. of IEEE/SICE International Symposium on System Integration (SII 2011)*, (2011), pp.1311-1316.

３．査読付き講演

[7] **山科亮太**，前原晋策，黄健，藪田哲郎,，“報酬変化を用いた Q-Learning による実ロボットの前進行動獲得”，第 10 回ロボティクスシンポジア予稿集，4D3, (2005), pp.411-417.

[8] 本山晴寿，**山科亮太**，黄健，藪田哲郎,，“強化学習を用いたロボットの前進行動形態に関する考察”，第 11 回ロボティクスシンポジア予稿集，(2006), pp.258-263.

4．一般講演

[9] **山科亮太**，林俊樹，藪田哲郎，“ニューラルネットワークを用いた 1 自由度フレキシブルアームの学習制御”，第 21 回日本ロボット学会学術講演会講演概要集，(2003), 1C16.

[10] **山科亮太**，前原晋策，石川智弘，藪田哲郎，“報酬変化に基づく Q-Learning の収束性に関する検討”，ロボティクス・メカトロニクス講演会'04，(2004), 2A1-L1-26.

[11] **山科亮太**，井上将志，浦川真理子，黄健，藪田哲郎，“報酬変化を用いた Q-Learning によるロボットの行動獲得”，ロボット・メカトロニクス講演会'05，(2005), 1A1-S-058.

[12] 黒田将史，**山科亮太**，藪田哲郎，”主観報酬学習を用いたイモムシ型ロボットの行動獲得”，ロボティクス・メカトロニクス講演会'12, (2012), 1A1-D06.

5．著者参考論文

[13] 池野谷康司, **山科亮太**, 原正之, 藪田哲郎, “1 自由度フレキシブルアームの反復学習制御”, 計測自動制御学会論文集，Vol. 41, No. 1, (2005), pp.91-93.

[14] 醒井雅裕，石ヶ谷康功，**山科亮太**，和井田匠，“定着制御のモデルベース開発に向けたシミュレーション技術の構築”，Imaging Conference Japan fall meeting 2010, (2010), pp.17-20.

[15] K. Nakamura, **R. Yamashina**, Shaobo Li, T. Kawabe: A Configuration of Model Predictive PID Control for Heat Conduction System, Proc. of SICE Annual Conference 2012, (2012), pp.187-191.

[16] 石井賢治，長藤秀夫，吉川政昭，**山科亮太**，“カラーQSU 技術（DH 定着方式）”，Ricoh Technical Report, No.38, (2012)，pp.44-49.

6．工業所有権

登録特許 13 件（国内 5 件 / 米国 8 件）

公開特許 30 件（国内 30 件）