東北大医保健学科紀要　15(1)：17～28，2006

価値関数を用いた非線形手動制御系の学習過程に関する考察

後藤太邦[1]，本間経康[2]，吉澤　誠[3]，阿部健一[4]

[1]東北大学大学院工学研究科　電気・通信工学専攻
[2]東北大学医学部保健学科　放射線技術科学専攻
[3]東北大学情報シナジーセンター　先端情報技術研究部
[4]日本大学工学部　情報工学科

The Value Function Analysis of the Learning Process on the Manual Control of Nonlinear System

Takakuni GOTO[1], Noriyasu HOMMA[2], Makoto YOSHIZAWA[3] and Kenichi ABE[4]

[1]*Department of Electrical and Communication Engineering Graduate School of Engineering, Tohoku University*
[2]*Department of Radiological Technology, School of Health Sciences Faculty of Medicine, Tohoku University*
[3]*Research Division of Advanced Information Technology Information Synergy Center, Tohoku University*
[4]*Department of Computer Science College of Engineering, Nihon University*

Key words: Manual Control, Nonholonomic System, Reinforcement Learning, Value Function

This paper describes analysis on human operator's trial and error learning process to control a nonholonomic system. A novel analysis technique using the value function of reinforcement learning is proposed. According to the transition of the value function, human operators tend to explore an objective trajectory first, and then shift to the following control of the trajectory and accelerate it. This acceleration disturbs the objective trajectory, and inducts another exploration phase to converge on the better solution. These results of the stepped approach of human learning may inspire an improvement to reinforcement learning.

1. はじめに

自動車におけるドライバーや，クレーンにおける操縦者などのように，人間は，機械と融合した人間－機械系を構成し，連続時間コントローラとしての役割を果たすことができる。このような手動制御系における人間オペレータの制御特性を解明することは，制御しやすい機械の設計や学習支援システムの構築，さらにはシステムの自動化に役立つことが期待されている。これらの目的のため，従来様々な研究が行われてきた。特に，Tustin [1] によって視覚入力と操作ハンドルの変位の関係が近似的に伝達関数で記述できることが示されて以来，主に制御工学の視点から解析する手法が用いられ，線形制御系に対する人間の特性に関しては多くの知見が得られている。しかしながら，人

間は線形制御系にとどまらず，複雑で強い非線形要素をもつ機械システムの操作も可能である。しかも制御対象の動特性に関する知識が無い状態から，入出力信号をもとに両者について何らかのモデルを試行錯誤的に学習し，適切な制御則を見出すことができる。人間の制御特性や学習特性をより深く調査するためには，制御対象を実システムによく見られる非線形システムへと拡張することが重要であると考えられる。強い非線形要素をもつ制御対象のひとつに非ホロノミック系 [2] がある。非ホロノミック系とは，拘束条件が積分不可能な微分方程式で記述される系のことをいう。非ホロノミック系の例として自動車をあげる。自動車への入力要素はアクセル，ブレーキによる車体速度とハンドルによる操舵角の2つである。車体の進行方向を基準とした場合，入力による制限のため，車体は左右に直接行くことは出来ない。しかしながら，例えば一旦前に進んだ後ハンドルを切り返してバックすれば結果として目標状態にたどり着くことが出来る。このように，非ホロノミック系の制御では，目標値にたどり着くまでの軌道を任意に選ぶことが出来ず，適切な軌道の計画と，計画した軌道をたどる追従制御の2つの混合問題を解かなければならない。制御理論の分野では，非ホロノミック系の制御に対する研究が盛んに行われている。例をふたつ挙げると，ひとつは Chained form という正準系への変換 [3] を行うことによって不連続フィードバック制御器を構成する方法，もうひとつは，2次形式で与えられる評価関数を最小にする入力を求める非線形最適レギュレータ問題に帰着させて解く方法 [4] がある。しかしながら系によっては Chained form への変換が不可能であるものも存在し，非線形最適レギュレータを構成する手法においては，評価関数の最小値である値関数を，Hamilton-Jacobi-Bellman (HJB) 偏微分方程式から求める必要があり，大抵は解析的に求めることができない。そのため，動的計画法などの数値探索が必要となる。このように，制御理論による取り扱いが困難な非ホロノミック系において，人間の制御動作を調査し，訓練によって良好な制御動作を獲得できたという報告がある。猪岡らは [5]，非ホロノミック系の単純な例である，第一関節が自由間接の2リンク劣駆動マニピュレータ（2PUAM : 2-Link Planer Underactuated Manipulator）の手動制御実験を行い，被験者の制御動作を応用した逐次的な制御則を提案した。谷貝らは [6] 第2関節を自由関節とした2PUAMについて猪岡らと同様の実験を行い，時間反転及び時間軸伸縮を用いた双方向アプローチによる 軌道計画と軌道追従制御則を提案している。しかし，これらの研究では人間がどのように試行錯誤し，制御動作を改善しているかについては観察的な記述にとどまり，オペレータが学習中に目標値までの経路をどのように計画し，また計画した軌道上の制御動作をどのように変容させているかなどの学習過程における詳細な報告はされていない。一方末長 [7] は，追跡手動制御系における人間の制御動作が，目標値特性に対する学習の進行度によって補償から追跡，予測動作へと変容するとした Successive Organization of Perception (SOP) 過程 [8] に着目し，試行の繰り返し中に目標値の時間軌道の形状を把握することが予測動作へ移行のきっかけになるという可能性を実験結果より示した。さらに同氏は，Norman が提唱する行為遂行の7段階理論のうちの，「意図の形成」および「行為の決定」に着目し，予測手動制御系における視覚支援情報がこれらの習熟過程に寄与するものと考えている [9]。双方の研究において，「目標値形状の把握」と「意図の形成」，及び「予測動作」と「行為の決定」は同義であると考えられる。ここでは自動車を制御対象とした手動制御実験を行い，「目標地点までの空間的軌道」と「制御対象の予測軌道」をそれぞれ「意図の形成」および「行為の決定」の視覚支援情報とした場合の訓練者の技能習熟特性について調べている。この結果から，被験者は目標空間軌道から意図の形成に関する方策を習熟した後，予測軌道の活用から行為の決定に関する方策を取得するといった段階的な学習を行っていることが示されている。しかしながら，教示の無い試行錯誤による場合に関する詳細な考察はされていない。上記の場合においても段階的な手順が踏まれる，と

いう可能性について検討するためには，人間が何に関して試行錯誤しているのか，また学習の進行状況に応じて制御動作のどの部分が改善されているのかをを何かしらの方法で検出する必要がある。ところで，未知の環境において最適な制御側を試行錯誤によって獲得するヒューリスティックな手法のひとつに強化学習［10］がある。強化学習は，観測される状態によって一意に決まる，報酬と呼ばれる強化信号の期待総和である価値関数をメモリ上に保持している。価値関数は各状態の「良さ」を示すものであり，価値関数を最大化する政策をみつけることで漸化的に目標行動を獲得する枠組みである。この価値関数の学習方式はTemporal Difference Learning（TD学習）［10］に代表されるように逐次的なものであり，ある方策に従った行動をとる前後関係を利用して更新される。つまり，ある行動に対する評価がその行動をとった直後に行われるため，学習開始時からの行動履歴に対する評価値が価値関数として記録されることに相当すると考えられる。本来強化学習アルゴリズムは行動決定も内部で行われる。しかし，行動の決定を仮に外部で行った場合，価値関数は外部方策を報酬と言う固定された基準によって評価する評価器と考えることができる。そこで本研究では，非ホロノミック系における制御問題を非線形レギュレータ問題として捉えた手動制御実験を行い，人間オペレータの学習過程を，目標軌道の探索とそれに伴った軌道追従制御の制御動作の変容の視点から検討することを目的とする。実験環境は谷貝ら［6］が用いた2PUAMと同様の環境を用いる。学習過程の解析には，評価関数で記述される制御成績の学習曲線，学習被験者がたどった軌道の解析，そして人間オペレータからの制御入力を強化学習における方策としたTD学習によって価値関数を記録する方法を提案し，学習の進行度に応じた上記3つの要素の変化を調べる。

2. リンク平面型劣駆動マニピュレータ

本研究で用いる2PUAMの概要について説明する。2PUAMはFig. 1に見られるように第2関節が非駆動関節となっている。座標系に関節角度座標を用いると，運動方程式は，

$$\mathbf{M}_{11}(\theta)\ddot{\theta}_1+\mathbf{M}_{12}(\theta)\ddot{\theta}_2+\mathbf{c}_1(\theta,\dot{\theta})=\gamma, \qquad (1)$$
$$\mathbf{M}_{21}(\theta)\ddot{\theta}_1+\mathbf{M}_{22}(\theta)\ddot{\theta}_2+\mathbf{c}_2(\theta,\dot{\theta})=0, \qquad (2)$$

と表せる。θ_1, θ_2 はそれぞれの関節角度，$\mathbf{M}$ は慣性行列，$\mathbf{c}$ はコリオリ・遠心力項を表す，そして各式の右辺は入力トルクを表す。式(2)が拘束条件となる。この系は，特殊な関節配置を除き，拘束条

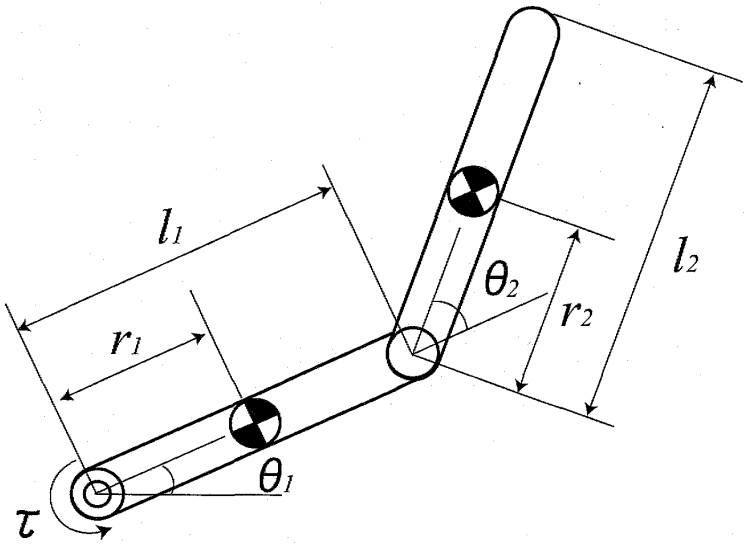

Fig. 1. 2PUAM.

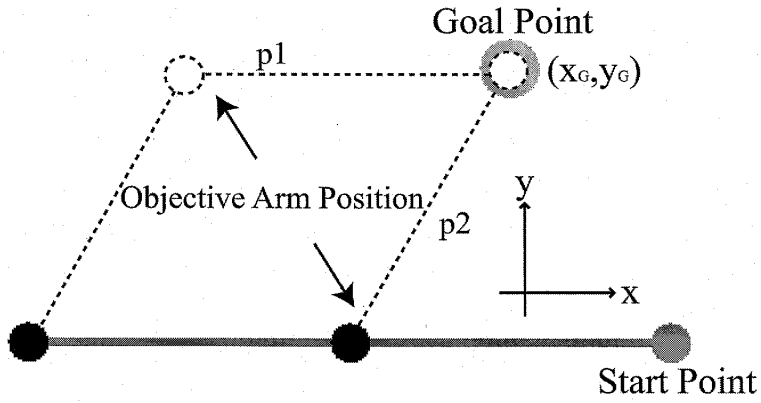

Fig. 2. Environment of Learning.

件が加速度まで含む積分不可能な微分方程式で表現される2階の非ホロノミック系である。このマニピュレータはすべての関節角が停止可能な平衡点であるが，平衡点近傍で線形化した系が可制御ではなく，線形コントローラの構築が不可能な系である［11］。

2.1 実験環境

Fig. 2 に 2PUAM を用いた制御タスクを示す。実験環境は，パーソナルコンピュータ上にサンプル時間 $\Delta t=0.02$(s) のルンゲクッタ法によって実装された，実時間シミュレーション環境を用いている。被験者は，ディスプレイ上に表示された 2PUAM を見ながらジョイスティックを用いて第1関節のトルク $\tau\in[-1,1]$ を操作する。制御目的はアーム先端部をスタート地点 $(\theta_1,\theta_2)=(0,0)$ から動かし，ゴール地点にて再び停止させることとする。ゴール地点は角度ではなく xy 平面の座標で与えられているため，図中の $p_1({\theta_1}^{p_1},{\theta_2}^{p_1})=(0,\pi/3)$, $p_2({\theta_1}^{p_2},{\theta_2}^{p_2})=(\pi/3,-\pi/3)$ の2通りの目標関節角をとることができる。制御成績の評価方法は，時刻 $k\Delta t$ $(k=0,1,2,\ldots,T)$ における作業空間上のアーム先端位置 $(x(k),y(k))$ とゴール位置 (x_G, y_G) の偏差 $\varepsilon(k)=\sqrt{(x(k)-x_G)^2+(y(k)-y_G)^2}$ を用いて，評価関数を

$$J=\sum_{k=0}^{T}\varepsilon(k) \qquad (3)$$

と定義し，各試行ごとに求める。T は試行時間を表す。以上の環境において，実験は 2PUAM に関する知識を持たない健常な20代男子4人を被験者として制限時間30秒，つまり $T=1500$ サンプルの試行を1試行，1セットを10試行とし，1日5セットずつ4日間，合計200試行を行った。また，被験者には，各試行において提示される評価値 J をできるだけ小さくするよう指示した。

3. 強化学習

各時点 k $(\in 0,1,2,\ldots)$ において，環境状態が $s_k=s\in S$ のとき，その状態観測に基づき，エージェントが行動 $a_k=a\in A(s_k)$ をとったとすると，次に可能な各状態 $s_{k+1}=s'$ への遷移確率は ${\mathcal{P}}^{a}_{ss'}=\Pr\{s_{k+1}=s'|s_k=s,a_k=a\}$ で与えられ，得られる報酬 $\gamma_{k+1}\in R$ の期待値は，${\mathcal{R}}^{a}_{ss'}=E\{r_{k+1}|s_k=s,a_k=a\}$ となる。上記のように遷移確率と報酬期待値が記述できる有限離散マルコフ過程において，方策 π に従うエージェントの，各状態においてそこから後に見込める報酬の割引期待総和である状態価値関数 $V^{\pi}(s)$ は

$$V^{\pi}(s)=E_{\pi}\{\sum_{j=0}^{\infty}\gamma^j r_{k+j+1}|s_k=s\} \qquad (4)$$

また，状態行動対に関する価値関数である，行動価値関数 $Q^{\pi}(s,a)$ は，

$$Q^{\pi}(s,a)=E_{\pi}\{\sum_{j=0}^{\infty}\gamma^j r_{k+j+1}|s_k=s,a_{k=a}\} \qquad (5)$$

で与えられる。E は期待値，$\gamma\in[0,1]$ は割引率を表わす。エージェントの目標は，この価値関数を最大化する方策 π^* を獲得することである。環境

と対峙したエージェントは試行錯誤を繰り返しながら，割引期待利得の最大化を目的として，各時点で得られる報酬 r_k に基づいて価値関数を更新していく。価値関数の更新アルゴリズムは，基本的に逐次更新型の TD 学習 [10] が用いられる。エージェントの方策はこの価値関数を用いて改善される。方策の改善アルゴリズムとしては，状態価値関数を用いた Actor-Critic[12]，行動価値関数を用いた Q-学習 [13]，Sarsa [14] などが広く知られている。しかしながら，本研究では行動の選択を人間が行うため，方策の改善は行わず，被験者が取る方策についての評価のみを行う。そこで今回は，人間の方策評価に状態価値関数を用いた TD(λ) [10] を適用する。2PUAM 環境における状態価値関数 V の表現方法及び TD(λ) のアルゴリズムを以下で説明する。

3.1 タイルコーディング

2PUAM の実験環境では連続な状態を扱うため，タイルコーディングを用いて状態価値関数の近似を行う。タイルコーディングは線形アーキテクチャの一つで，二次元状態空間を例に取る Fig. 3 のように状態を分割するタイリングを複数枚重ねたものである。各タイリングの要素はタイルと呼ばれ，その１つ１つにそれぞれ特徴量ベクトルの要素 ϕ 割り当てられる。参照された状態 s が含まれる領域，つまり s が含まれる各タイリングのタイル内の領域を特徴集合 $F(s)$ とすると，$F(s)$ に該当するタイルの特徴量ベクトルの値は 1 に，それ以外は 0 となる。タイルコーディングを用いて価値関数 V を表現する場合，特徴量ベクトルと同数のパラメータ v を用いて，

$$\phi_i(j)=\begin{cases}1 & \text{for } i,j\in F(s)\\ 0 & \text{for } i,j\notin F(s)\end{cases} \tag{6}$$

$$V(s)=\sum_{i,j} v_i(j)\phi_i(j) \tag{7}$$

と表される。ここで $i\,(i=1,2,\ldots,n)$ はタイリングの番号，$j\,(j=1,2,\ldots,m)$ は各タイリングにおけるタイルの番号を表す。

3.2 TD (λ)

連続状態における TD(λ) では，遷移後の状態 s' における V を評価値として用いる。ここでは連続状態を取扱うので，式(7)における V のパラメータ v を更新する。

$$\eta_i(j)\leftarrow\begin{cases}1 & \text{for } i,j\in F(s)\\ \gamma\lambda\eta_i(j) & \text{for } i,j\notin F(s)\end{cases} \tag{8}$$

$$\delta=r+\gamma V(s')-V(s) \tag{9}$$

for all i and j

$$v_i(j)\leftarrow v_i(j)+\alpha\delta\eta_i(j) \tag{10}$$

$\alpha\,(0<\alpha\leq 1)$ は学習率を，$\gamma\,(0\leq\gamma\leq 1)$ は割引率を表す。また η は replacing eligibility trace と呼ば

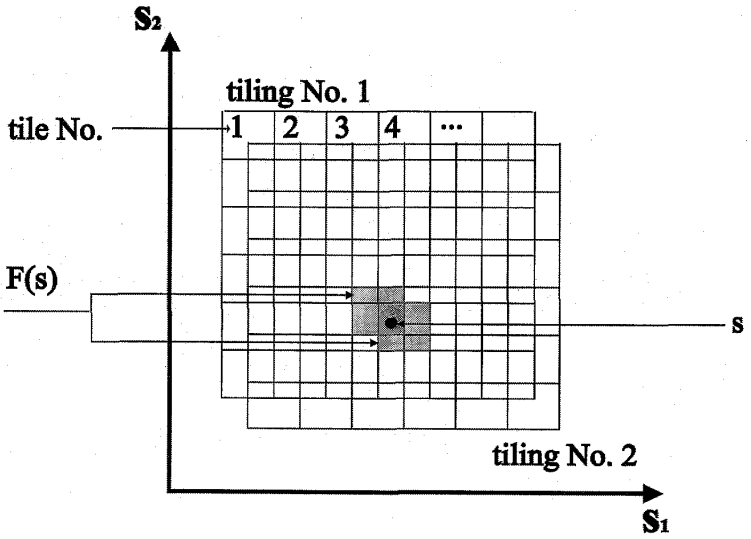

Fig. 3. Tile Coding.

れ，λ は $0 \leq \lambda \leq 1$ なる実数である。eligibility の計算は V の更新に先立って行われる。

3.3 実験で用いるパラメータ

実験環境において，各関節角及び角速度 θ_1, θ_2, $\dot{\theta}_1$, $\dot{\theta}_2$ の値を状態として観測する。状態は $21 \times 21 \times 1 \times 11$ のタイルに分割されたタイリングを 10 枚重ねて表現した。V の更新には TD(λ) を用い，$\alpha = 1/m$ (m はタイリングの枚数)，$\lambda = 0.9$，$\gamma = 0.99$, 価値関数パラメータの初期値を $v_0 = 0$ とした。報酬は，評価関数 J を基準に報酬関数

$$r(k) = \exp\left(-\frac{\varepsilon(k)}{2\delta}\right) \qquad (11)$$

を設計した。価値関数の更新は各時点にて行った。

4. 実験結果と考察

4.1 評価値の推移と制御動作の変容

Fig. 4 に評価値の推移曲線を示す。横軸が試行回数，縦軸が式(3) で表される J を表す。結果を見ると，まず全般的に初期の段階で大きな変動が続き，その後試行が進むにつれて J が小さくなっている。このことから，被験者は試行を繰り返すことによって，何らかの方法で与えられた評価基準に対する制御動作を改善できていることが確認できる。

次に，各被験者の推移形状について見ると，被験者 A の評価値はほぼ単調に減少しているのに対し，他の被験者の評価値は一旦収束する兆候を見せた後，図中破線領域内で再び値に変動が生じている。特に被験者 D の場合は，破線領域に入る前の最小評価値 J_{min} の値が $J_{min} = 0.0105$ であったのに対し，破線領域を超えた後では $J_{min} = 0.0049$ と小さくなっていた。被験者 C は評価値の変動が収まる前に実験が終了してしまったが，追加実験を行ったところ，被験者 D と同様な結果が得られた。

具体的に被験者がたどっていた軌道例を Fig. 5 に示す。被験者 A, B 及び被験者 C, D の Fig. 4 における破線領域前の制御動作は，図中の Type 1 に見られるように，第 2 関節の角度を目標値近くに移動させ，その後第 2 関節を動かさないようにしながら第 1 関節を目標値に近づける操作を実行していた。アームの先端の動きを見てみると，デカルト座標系におけるゴールとの偏差がほぼ単調に減少している。しかし，この軌道では目標値付近で第 2 関節を減速できずに通過してしまう。そ

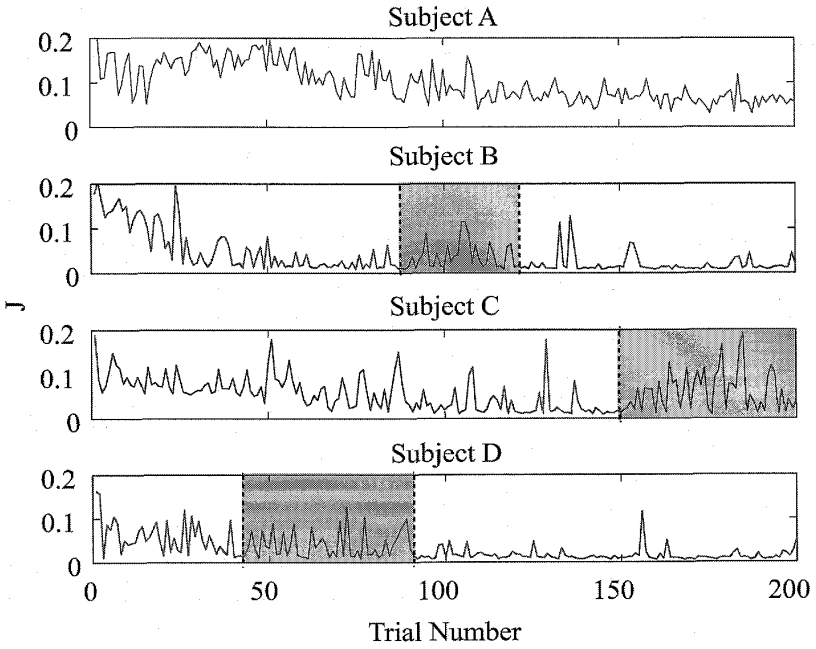

Fig. 4. Transition of evaluation value.

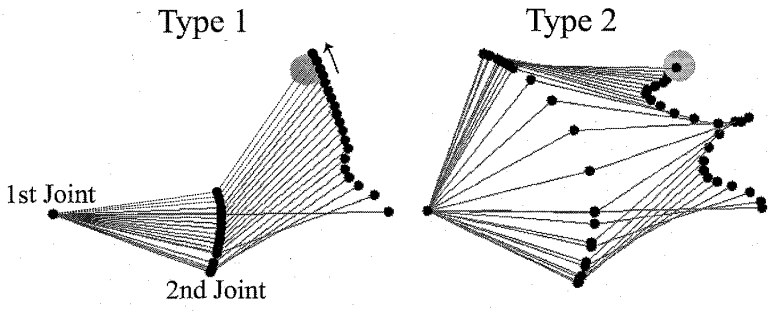

Fig. 5. Motions of 2PUAM controled by Subjects.

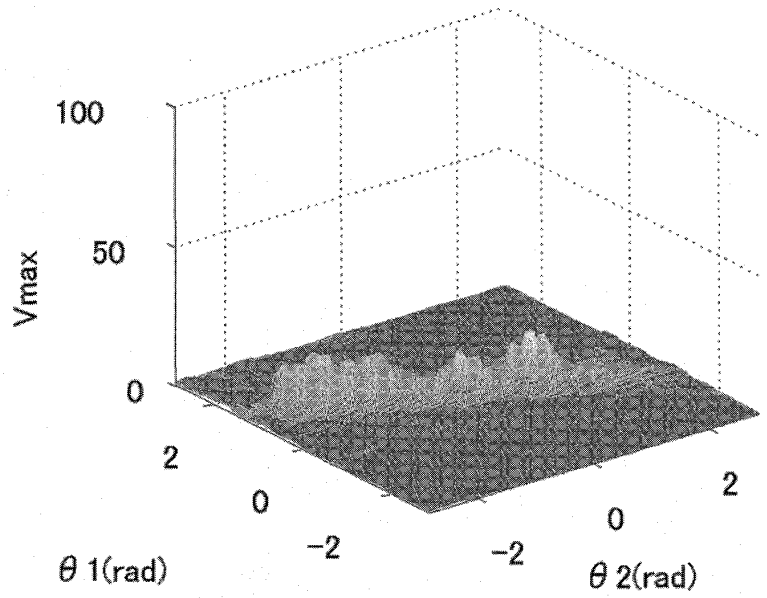

Fig. 6. The Value function map after 30 Trials.

のためA以外の被験者は速度を十分に落としてからゆっくり目標値へ近づくよう，非常に小さい入力を行っていた。一方，破線領域後の被験者C, Dの制御動作はType 2のような制御動作に変化していた。Type 2の軌道は目標値付近で両関節を同時に減速させることができ，うまく操作すれば目標値付近でほぼ停止させることが可能である。

4.2 価値関数の推移

被験者B, C, Dにおける評価値の推移曲線において，値の大きな変動がしばらく続いた後収束に向かい，再び大きな変動を迎えるという傾向が見られた。また，値の変動後に被験者CやDのたどった軌道の形状に変化が見られた。これらの結果から，変動の大きな場所と収束に向かっている場所において，それぞれ性質の異なる学習を行っていると考えられる。そこで，変動が起きる境界領域における被験者の行動について調べるため，変動が起きる直前，変動が収まる直前，変動が終了した後のそれぞれの時点における価値関数の解析を行う。解析は紙面の都合上最も特徴的な学習過程を見せていた被験者Dを代表として行う。

Fig. 6に最初の変動が収まる直前の30試行目における価値関数を，Fig. 7にその等高線を示す。価値関数$V(s)$における状態sは4次元である。そのためx軸，y軸を各関節の角度とし，ある位置の価値関数の表示値を，その位置における角速度の領域内での最大値 $V_{max}(\theta_1, \theta_2)=\max_{\dot{\theta}_1,\dot{\theta}_2}\hat{V}(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2)$ としている。Fig. 7の矢印は $\dot{V}_{max}$ をも

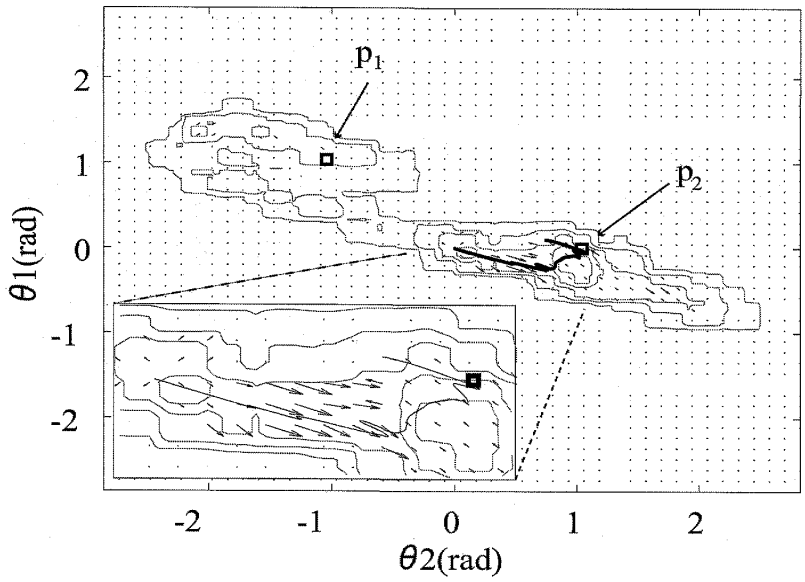

Fig. 7. The Contour Plot of The Value Function after 30 Trials.

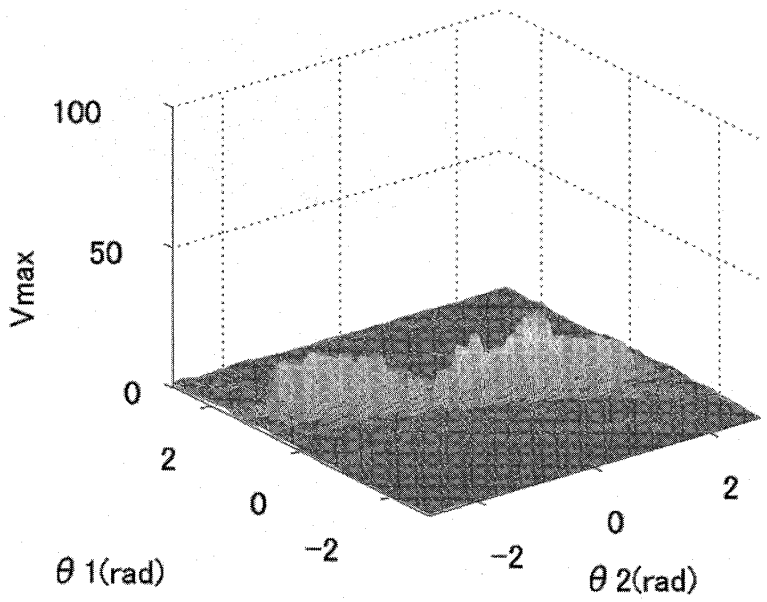

Fig. 8. The Value function map of after 40 Trials.

つ角速度ベクトルであり，矢印の長さは速度の大きさを表す。つまり，位置の重複が無ければ，ベクトルの方向にそった軌道が最良の軌道であるといえる。ベクトルは価値関数の最大値を基準とした閾値 V_{th} 以上の位置のみ表示している。また，□は各目標値の角度座標を表している。実際にそれぞれの図を見比べると，価値関数のピークが Fig. 2 で説明した目標関節角のひとつである p_2 の辺りにできているのが見れる。このことから，被験者は p_2 までの空間的な軌道を形成できているものと思われる。Fig. 7 の黒線は実際に 30 試行辺りで被験者がたどった軌道であるが，この軌道の形状と，ベクトルをなぞってできる軌道は類似したものとなっている。次に，2 回目の変動が起きる直

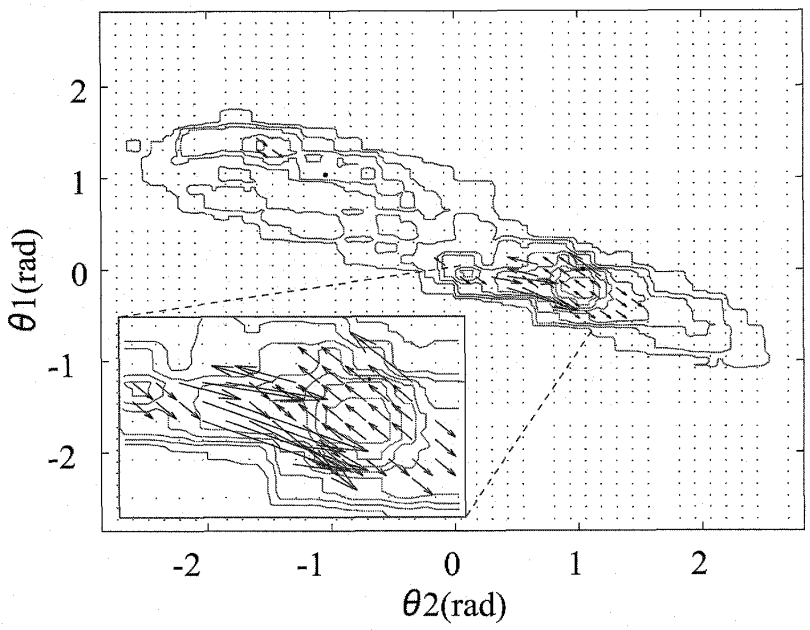

Fig. 9. The Contour Plot of The Value Function after 40 Trials.

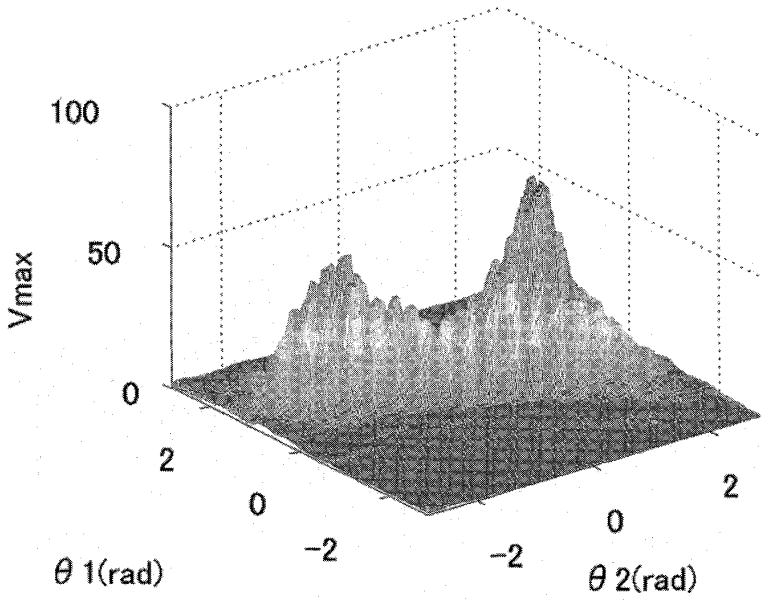

Fig. 10. The Value function map after 90 Trials.

前の40試行目における価値関数をFig. 8及びFig. 9に示す。価値関数の形状は，あまり試行回数が増えていないこともあり，p_2のピークが少し高くなっている程度である。一方，Fig. 9を見てみると，ベクトルの向きが30試行目のときとあまり変わらないのに対し，ベクトルの長さが全般的に長くなっている。このグラフから，被験者は軌道の空間的な形状を保ちながら，通過速度を高速化する試みを行っているといえる。ここで，Type 1の軌道は，p_2付近での減速が不可能である。そのため，軌道の通過を高速化するほど，高速でp_2を通り過ぎてしまい，不安定な状態を招きやすくなる。

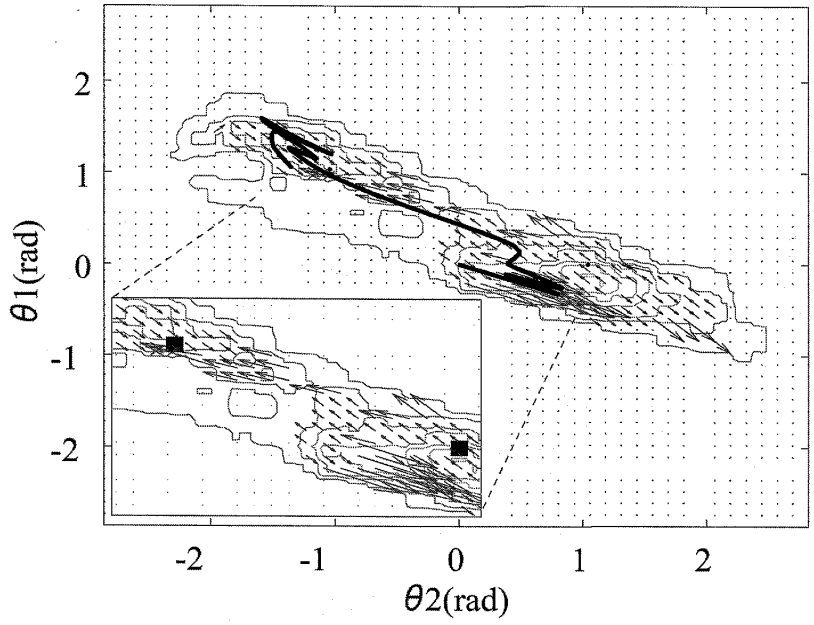

Fig. 11. The Contour Plot of The Value Function after 90 Trials.

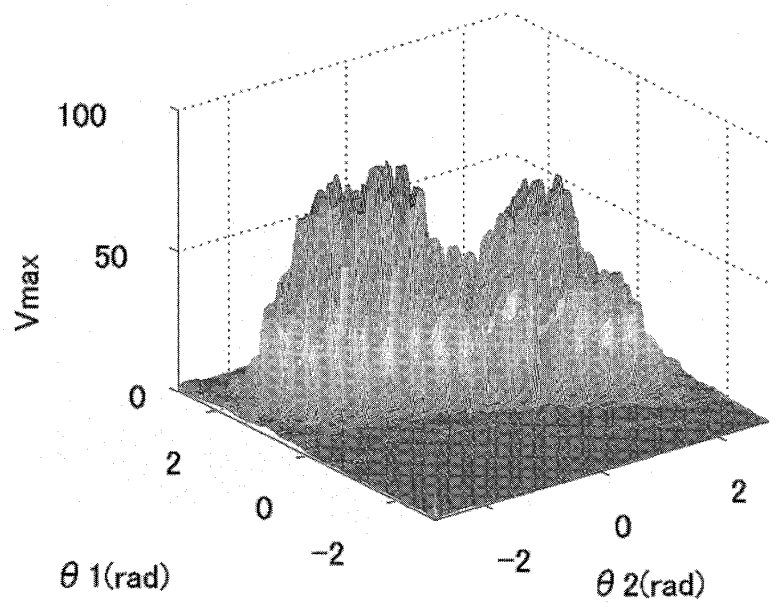

Fig. 12. The Value function map after 200 Trials.

この不安定な状態が，次におこる変動の発生の原因になっていると考えられる。2回目の変動が収まる直前の90試行目における価値関数をFig. 10，及びFig. 11に示す。Fig. 10をみると，p_1の位置にピークが出てきているのがわかる。またFig. 11に見られるように，ベクトルの矢印及び被験者のたどった軌道がp_1に伸びている。Fig. 12，及びFig. 13は実験終了後の価値関数である。90試行目と比べると，明らかに矢印の長さが伸びている他に，p_1におけるピーク値のほうがp_2のそれと比べて大きくなっている。これは軌道通過速度の向上と言う点で，30試行目から40試行目にかけて見られた被験者の挙動と酷似している。以上の実験結果より，被験者の学習過程において主に3つ

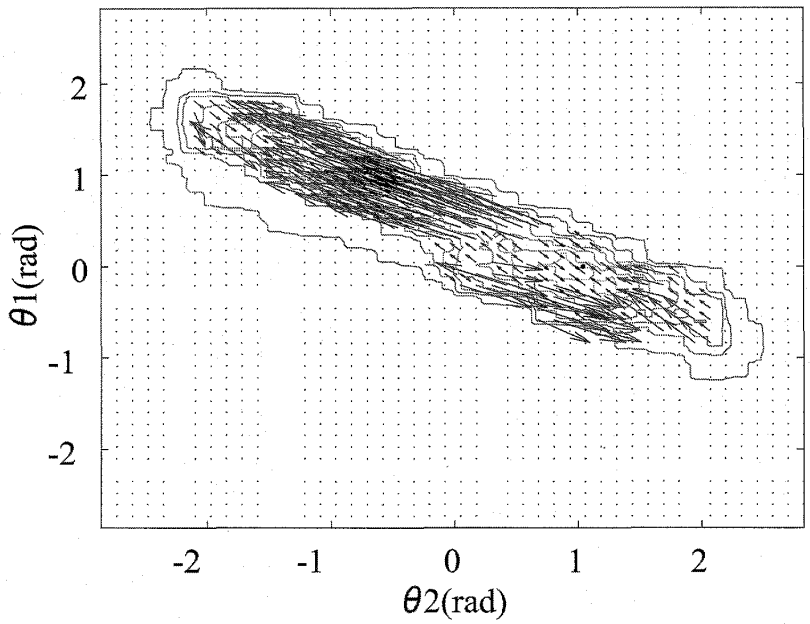

Fig. 13. The Contour Plot of The Value Function after 200 Trials.

の特徴が挙げられる。

- 評価値の変動が収まり始める辺りまでに，目標値までの空間軌道を形成
- 変動が収束に向かっている最中では，形成された目標空間軌道の形状はあまり変化せず，軌道を通過する速度が向上
- 高速化による軌道の不安定化によって次の変動が発生

これらの特徴より，少なくとも本実験環境における被験者の学習過程は，末長によって提唱されている「意図の形成」および「行為の決定」の段階的な手順が教示の無い試行錯誤による学習過程においても踏まれていることを強く支持する結果となった。これに加え，行為の決定においては意図，つまり目標軌道への追従のみならず，軌道通過の高速化も行われていることがわかった。高速化によって目標軌道への追従が不安定化されることにより，「意図の形成」に関する学習が再び誘発されることから，非ホロノミック系における手動制御の学習過程は段階的な手順を繰り返す逐次的なものである可能性を示す結果が得られたと考えられる。

5. ま と め

本論文では，非ホロノミック系における人間オペレータの試行錯誤による学習過程を，目標軌道の探索と目標軌道への追従制御動作の変容の視点から検討した。2PUAM を制御課題とした手動制御実験を行い，人間の行動履歴を価値関数によって評価した。価値関数の解析結果より，制御課題に対する人間の学習過程は，目標軌道の探索→目標軌道への追従と通過速度の向上という逐次的な構造を持っていることが示された。各段階において人間は，試行錯誤を行う領域を限定していると考えることができる。あえて探索領域を制限することで学習の効率化を図っているのならば，強化学習における探索手法にも有効である可能性がある。今後は本実験結果の一般性を別の環境にて検証することとともに。本実験結果から得られた知見をもとにした，強化学習をベースとした人間の段階的学習アルゴリズムの考案が今後の課題である。

文　献

1) Tustin, A.: The nature of the operator's response in manual control and its implication for controller design, J.I.E.E., 94-2A, 1947
2) 中村仁彦：非ホロノミック系制御研究の展望，計測と制御，**36**, 384-389, 1997
3) Astolfi, A.: Discontinuous control of nonholonomic systems, Systems. and Contro Letters, **27**, 37-45, 1996
4) Imafuku, K., Yamashita, Y., Nishitani, H.: Control of a wheeled vehicle using the viscosity solution of the Hamilton-Jacobi partial differential equation, JRS J, **17**, 689-695, 1999
5) Inooka, H., Shito, Y., Yu, K.: Manual control of the two-link arm with a free joint, Prc. of IEEE International Conference on Systems, Man and Cybanetics, 2324-2328, Oct. 22-25, 1995
6) Yagai, M., Ishihara, T., Inooka, H.: Manual control of the positioning of two-link arm with a free joint, SICE Tohoku chapter workshops, 2061, 2002
7) Suenaga, O.: A study on human operator's learning process in manual control systems —An analysis by shape of reference signal and control operation—, 日本人間工学会誌，**34**, 177-189, 1998
8) McRuer, D.T., Allen, R.W., Weir, D.H., Klein, R.H.: New results in driver steering control models, Human Factors, **19**, 381-397, 1977
9) Suenaga, O.: A basic analysis on skill learning characteristics by displaying visual aid information in manual controls, 日本人間工学会誌, **40**, 272-275, 2004
10) Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction, MIT Press, 1-322, 1998
11) Oriolo, G., Nakamura, Y.: Free-joint manipulators: Motion control under second-order nonholonomic constraints, Proc. of IROS, 1248-1253, 1991
12) Barto, A., Sutton, R., Anderson, C.: Neurolike adaptive elements that can solve difficult learning control problems, IEEE Trans. on Systems, Man, and Cybern etics, SMC13, 834-846, 1983
13) Watkins, C.J.C.H., Dayan, P.: Q-learning, Machine Learning, 8, 279-292, 1992
14) Rummery, G.A., Niranjan, M.: On-line Q-learning using connectionist systems, Technical Report CUED/F-INFENG/TR 166, Engineering Department, Cambridge University, 1994