計測自動制御学会東北支部 第225回研究集会 (2005.11.11)
資料番号 225-4

非ホロノミック系における手動制御の学習過程

A Leaning Process on the Manual Control of Nonholonomic Systems

○後藤太邦*，本間経康*，吉澤誠*，阿部健一**

○Takakuni Goto*, Noriyasu Homma*, Makoto Yoshizawa*, Kenichi Abe**

*東北大学, **日本大学

*Tohoku University, **Nihon University

キーワード： 手動制御 (manual control)，非ホロノミック系(nonholonomic systems)，強化学習 (reinforcement learning)，価値関数 (value function)

連絡先： 〒980-8579 仙台市青葉区荒巻字青葉6-6-05 東北大学大学院 工学研究科 電気・通信工学専攻 吉澤研究室
後藤太邦，Tel.: (022)795-7130，Fax.: (022)263-9163，E-mail: goto@yoshizawa.ecei.tohoku.ac.jp

1. はじめに

人間は，複雑で強い非線形要素をもつ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．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．このような非線形手動制御系における人間の学習特性を解明することは，学習支援システムや自律学習システムの開発に寄与するものとして期待できる．

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[1, 2, 3, 4]がある．非ホロノミック系と線形系の大きな違いは，目標値までの軌道を任意に選ぶことができない点にある．そのため，非ホロノミック系の制御では，目標値に対する偏差のフィードバックループに，目標値までの軌道の計画を，何かしらの形でおりまぜなければならない．非ホロノミック系の制御に関する研究[5, 6]は盛んに行われているが，上記の問題に対して改良の余地が残されている系も少なくない．

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，人間オペレータの制御動作を調査し，動特性が未知な状態から試行錯誤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[7]，改良の余地が残されている問題の一例である，第一関節が自由関節の2リンク劣駆動マニピュレータ(2PUAM:2-Link Planer Underactuated Manipulator)を制御対象とした手動制御実験を行い，訓練後のオペレータの制御特性を応用した逐次的な制御則を提案した．谷貝らは[8]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．但しこれらの報告は，学習後における人間の制御特性の調査に重きがおかれており，学習特性に関する詳細な報告はされていない．というのも，試行の繰り返し中に人間がどの部分で試行錯誤し，制御動作を改善しているのかを検出ことが困難であることが原因であると考えられる．

そこで本研究では，人間が試行錯誤し，制御則を改善した領域を調べる手段として，強化学習[10]の枠組みを用いた解析手法を提案する．強化学習の枠組みでは，環境の状態を変数とした,「価値関数」といわれる各状態の「良さ」を表す関数がある．この価値関数は，ある制御則にしたがってオンラインで形成され，良い制御則になるほど価値関数は大きい値になる．そのため，人間の行動を制御側とした価値関数の形成過程を図として表示すれば，価値関数の大きくなった場所，つまり人間の制御則の改善された部分を視覚的に検出できる可能性がある．このような仮定から提案した解析手法を用いて，非ホロノミック系の手動制御における人間オペレータの，教示が無い場合の学習過程を検討する．手動制御実験は谷貝ら[8]が用いた2PUAMと同様の環境を対象とする．

2. 強化学習

2.1 強化学習と人間の最適制御

未知の環境において最適な制御則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がある．強化学習における行動決定者(エージェント)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」を状態空間上に保持している．強化学習とは，各状態の価値関数を最大化する制御則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．この価値関数を最大化するという問題設定は，評価関数を最小化する最適制御の枠組みと酷似しており，価値関数と評価関数は双対な関係にあるといえる[9]．つまり，ある環境において人間が価値関数のような何かしらの評価基準をもち，それを良くするように人間が試行錯誤することは，強化学習の枠組みに近いといえる．このような人間の評価基準を知ることができれば，学習の進行に応じた評価基準の変化により，制御則のどの部分が変化したかを知ることができる．しかしながら，このような人間の評価基準を直接知ることはきわめて困難である．一方で，仮に予め決められた評価関数を人間に提示すれば，人間の評価基準を評価関数によって誘導することは可能であると考えられる．そこで人間に予め評価関数を提示して試行錯誤を行わせ，人間の行動から価値関数を形成すれば，価値関数の形成過程を調べることにより，人間の評価基準の変化を間接的に推察することができると考えられる．そこで本研究では，人間の行動から価値関数を形成する手法を提案する．次節以降に強化学習のアルゴリズム及び実装方法について説明する．

2.2 強化学習の基本構造

各時点$k(\in 0,1,2,\cdots)$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，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$r_{k+1} \in R$の期待値は，$\mathcal{R}^a_{ss'} = \Pr\{r_{k+1} | s_k = s, a_k = a\}$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，ある制御則πに従う場合の状態sにおける価値関数$\mathrm{V}^\pi(s)$とは，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

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

で定義される．Eは期待値，$\gamma \in [0,1)$は割引率を表わす．強化学習の目標は，この価値関数を最大化する最適制御則π^*を獲得することである．最適制御則π^*における価値関数$\mathrm{V}^*(s)$は，動的計画法では環境のモデル$\mathcal{P}^a_{ss'}$，$\mathcal{R}^a_{ss'}$を既知として，Bellman方程式

$$\begin{aligned} \mathrm{V}^*(s) &= \max_a E_{\pi^*}\{r_{k+1} \\ &\quad + \gamma \mathrm{V}^*(s_{k+1}) | s_k = s, a_k = a\} \quad (2) \\ &= \max_a \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma \mathrm{V}^*(s_{k+1})] \quad (3) \end{aligned}$$

を解く問題へと帰着される．一方，強化学習では環境のモデル及び価値関数の真値$\mathrm{V}^\pi(s)$が未知である．そのため試行錯誤を繰り返しながら，各時点で得られる報酬r_kに基づいて価値関数の推定値$\hat{\mathrm{V}}(s)$を真値に近づけるように更新していく．価値関数の更新アルゴリズムは，基本的にTemporal Difference Learning(TD学習)[10]が用いられている．制御則はこの価値関数を用いて改善される．改善アルゴリズムとしては，Actor-Critic[11]，Q-学習[12]，Sarsa[13]などが広く知られている．本研究では行動の決定を人間が担うため，制御則の改善は行わず，価値関数の更新のみを行う．価値関数の更新アルゴリズムであるTD学習を以下で説明する．

2.3 TD学習

式(1)を変形すると，

$$\mathrm{V}^\pi(s) = E_\pi\{r_{k+1} + \gamma \mathrm{V}^\pi(s_{k+1}) | s_k = s\} \quad (4)$$

のようになる．つまり制御則πにおいて，状態遷移前後の価値関数の間には上式の関係が成り立つ．しかしながら，強化学習では，各状態の価値関数の推定値を観測するため上式の関係に誤差δが生じる．δはTD(Temporal Difference)誤差と呼ばれ，ある時点kにおいて，$s_k = s$，$s_{k+1} = s'$，$r_{k+1} = r$とすると，δは

$$\delta = r + \gamma \hat{\mathrm{V}}(s') - \hat{\mathrm{V}}(s) \quad (5)$$

で与えられる．これを用いて状態s価値関数の推定値の更新を，

$$\hat{\mathrm{V}}(s) \leftarrow \hat{\mathrm{V}}(s) + \alpha\delta \quad (6)$$

によって行う．$\alpha(0 < \alpha \leq 1)$は学習率を表す．TD(0)と呼ばれるこの方法は，価値関数の更新に1時刻先の情報しか用いられず，学習に多くの更新回数を必要とする．一方，各状態に訪問した履歴ηを

$$\eta_{k+1}(s_i) \leftarrow \begin{cases} 1 & \text{for } s_i = s_k \\ \gamma\lambda\eta_k(s_i) & \text{for } s_i \neq s_k \end{cases} \quad (7)$$

によって各時点ごとに求め，以前訪問した状態の価値関数を

$$\hat{\mathrm{V}}(s_i) \leftarrow \hat{\mathrm{V}}(s_i) + \alpha\delta\eta(s_i) \quad (8)$$

のようにをまとめて更新し，学習の高速化を図る方法がある．λは$0 \leq \lambda \leq 1$なる実数である．これはTD(λ)と呼ばれる方法である．

2.4 タイルコーディング

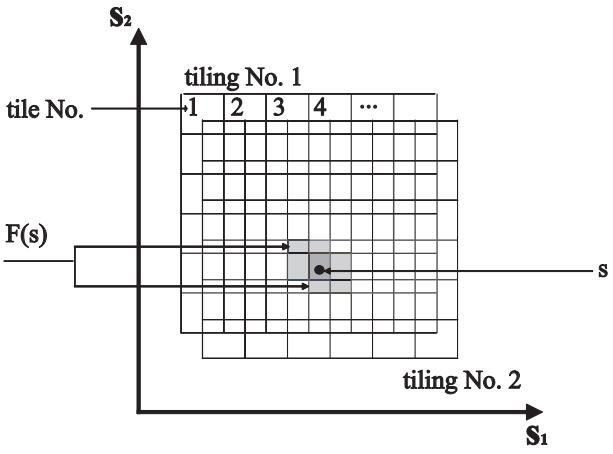

Fig. 1　タイルコーディング

前節までに説明した強化学習アルゴリズムは離散状態空間を対象としており，価値関数はテーブルルックアップ形式となっている．しかしながら，手動制御の対象である機械システムは状態が連続であるため，価値関数の近似が必要となる．本節

では，近似手法の一つであるタイルコーディングについて説明する．

図1のような2次元連続状態空間を例に取ると，状態を離散化する格子が2枚ずれながら重なっている．この格子のことをタイリングといい，格子のます目1つ1つのことをタイルという．各タイルは価値関数のパラメータを保有しており．i = 1,2,...,n番目のタイリングにおける，j = 1,2,...,m番目のタイルのパラメータは，$v_i(j)$と記述できる．ここで，図1のようにある状態の1点sが参照されると，価値関数$\hat{V}(s)$はsが含まれる領域，つまりsを含む各タイリング及びタイルの集合$F(s)$を用いて，

$$\hat{V}(s) = \sum_{i,j \in \mathrm{F}(s)} v_i(j) \tag{9}$$

と表される．

2.5 価値関数更新アルゴリズムの実験環境への実装

本研究では，人間の行動によって得られる状態遷移の軌跡を用いて価値関数を形成し，各試行後の価値関数の形状から，人間がとっている制御則の変化を推察する．価値関数はタイルコーディングによって表現し，TD(λ)を用いて更新する．以下に本研究で用いた価値関数更新アルゴリズムを示す．

全てのv,ηを0初期化
各試行に対して繰り返し:
　sを初期化，$k = 0$とする
　試行の各時刻に対して繰り返し:
　　$a \leftarrow$ 被験者によって入力された行動
　　行動aを取り，報酬rと次状態s'を観測
　　$\delta = r + \gamma \hat{\mathrm{V}}(s') - \hat{\mathrm{V}}(s)$
　　$\eta_i(j) \leftarrow \begin{cases} 1 & \text{for } i,j \in \mathrm{F}(s) \\ \gamma\lambda\eta_i(j) & \text{for } i,j \notin \mathrm{F}(s) \end{cases}$
　　$v_i(j) \leftarrow v_i(j) + \alpha\delta\eta_i(j)$
　　$s \leftarrow s'$
　　$k \leftarrow k + 1$
　　$k = T$ならば繰り返しを終了

3. 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．図2は2PUAMの略図である．図に見られるように，第2関節は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，

$$\mathbf{M}_{11}(\theta)\ddot{\theta_1} + \mathbf{M}_{12}(\theta)\ddot{\theta_2} + \mathbf{c}_1(\theta,\dot{\theta}) = \tau, \tag{10}$$

$$\mathbf{M}_{21}(\theta)\ddot{\theta_1} + \mathbf{M}_{22}(\theta)\ddot{\theta_2} + \mathbf{c}_2(\theta,\dot{\theta}) = 0, \tag{11}$$

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．また各式の右辺は入力トルクを表し，式(11)が入力トルク0という拘束条件となる．この系は，特殊な関節配置を除き，拘束条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[4]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[14]．

4. 実験環境及び実験方法

図3，4に実験環境及び制御タスクの略図を示す．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

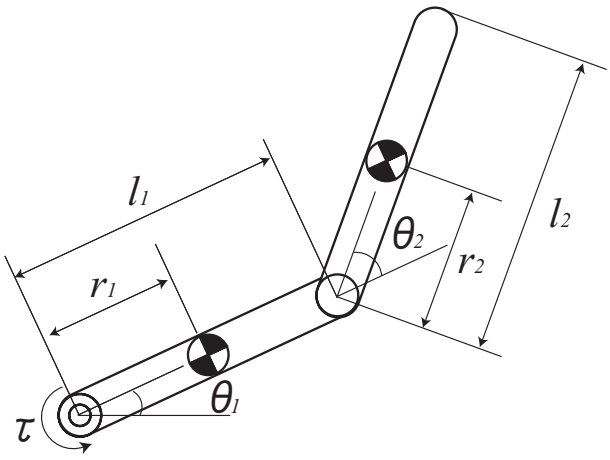

Fig. 2　2リンク平面型劣駆動マニピュレータ

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．制御目標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$k\Delta t(k = 0, 1, 2, \cdots, T)$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$\epsilon(k) = \sqrt{(x - x_G)^2 + (y - y_G)^2}$を用いて，評価関数を

$$J = \sum_{k=0}^{T} \epsilon(k) \tag{12}$$

と定義し，各試行ごとに求める．Tは試行時間(サンプル数)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．

4.1　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．状態はタイルコーディングを用い，$21 \times 21 \times 11 \times 11$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．$\hat{\mathrm{V}}$の更新にはTD(λ)を用い，$\alpha = 1/m$(mはタイリングの枚数), $\lambda = 0.9$, $\gamma = 0.99$とした．報酬は，評価関数Jを基準に報酬関数

$$r(k) = \exp(-\frac{\epsilon(k)}{2\sigma}) \tag{13}$$

を設計した．

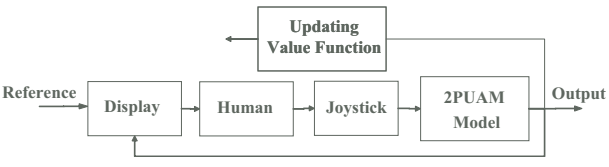

Fig. 3　実験システムのブロック図

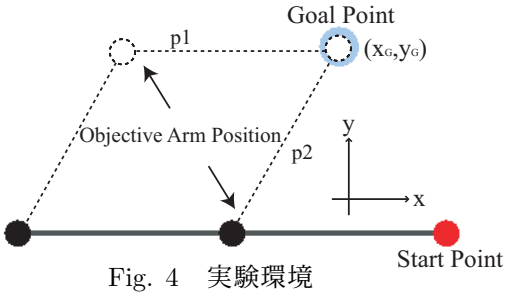

Fig. 4　実験環境

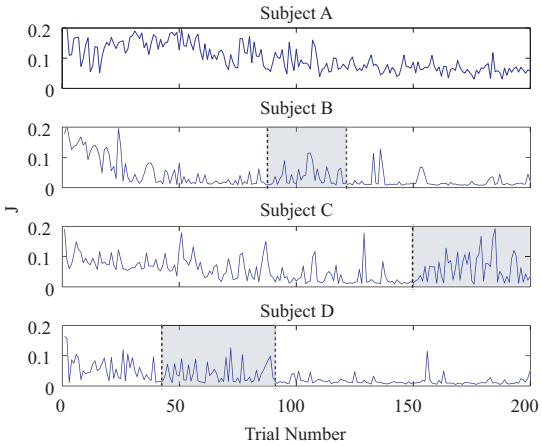

Fig. 5　評価値の変化の様子

5.　実験結果と考察

5.1　評価値の推移と制御動作の変容

図5に式(12)で表される評価値Jの推移曲線を示す．横軸が試行回数，縦軸が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

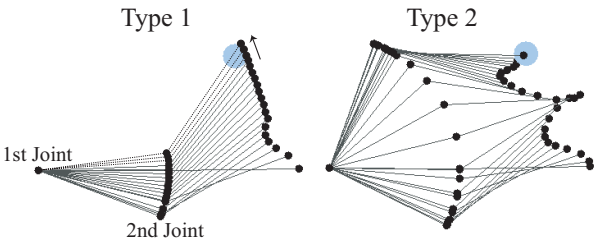

Fig. 6　被験者がとった2PUAMの軌道例

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．被験者Cの場合，評価値の変動が続いている時点で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B，Dと同様に最終的に変動は収束した．ここで被験者B，C，Dが破線領域に入る前と通過した後の評価値を比べると，全員値が小さくなっており，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$J_{min} = 0.0105$であったのに対して破線領域を通過した後では$J_{min} = 0.0049$と大幅に小さくなっていた．

具体的に被験者がたどっていた軌道例を図6に示す．被験者A，B及び被験者C，Dの図5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ゴールのデカルト座標系における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．

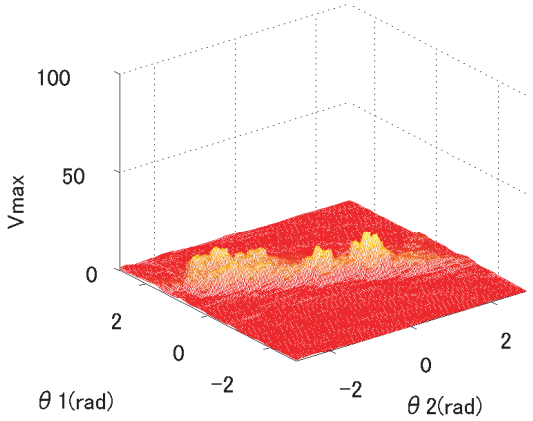

Fig. 7　30試行後の価値関数マップ

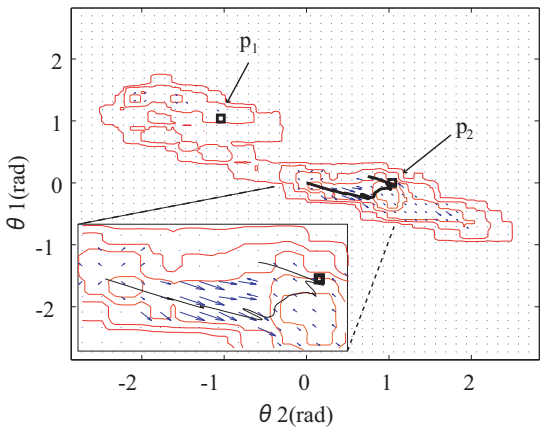

Fig. 8　30試行後の価値関数の等高線と各位置における各速度ベクトル

5.2　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，値の変動後に被験者CやDのたどった軌道に変化が見られた．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，性質の異なる学習を行っていると考えられる．そこで次に，評価値の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

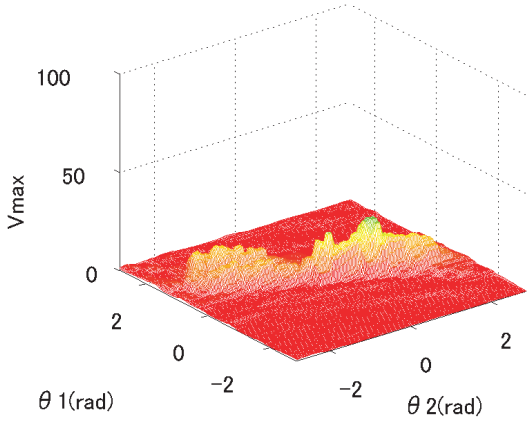

Fig. 9　40試行後の価値関数マップ

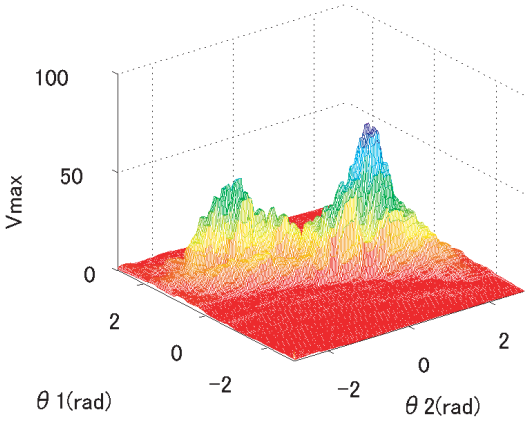

Fig. 11　90試行後の価値関数マップ

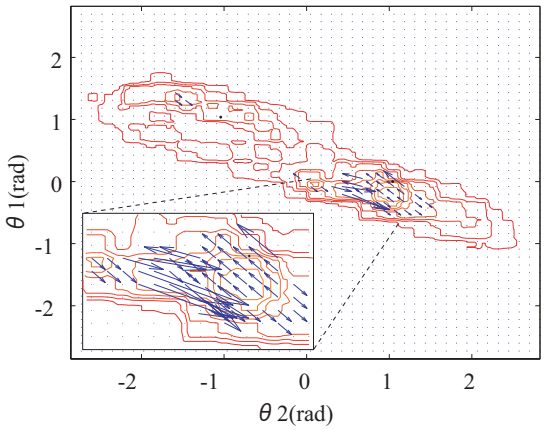

Fig. 10　40試行後の価値関数の等高線と各位置における各速度ベクトル

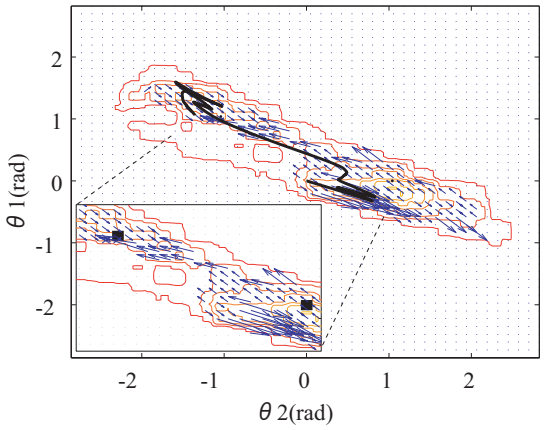

Fig. 12　90試行後の価値関数の等高線と各位置における各速度ベクトル

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．図7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，図8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$\mathrm{V}_{max}(\theta_1,\theta_2)=\max_{\dot{\theta}}\hat{\mathrm{V}}^{\dot{\theta}}(\theta_1,\theta_2)$としている．図8の矢印は$\mathrm{V}_{max}$をも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適な軌道であるといえる．

角速度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$p2$までの空間的な軌道を形成できているものと思われる．図8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前の40試行目における価値関数を図9及び図10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，図10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

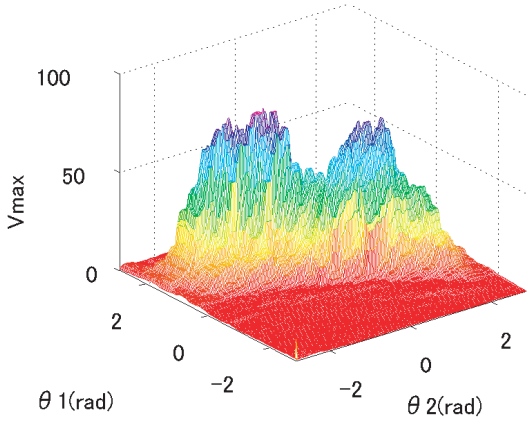

Fig. 13　200試行後の価値関数マップ

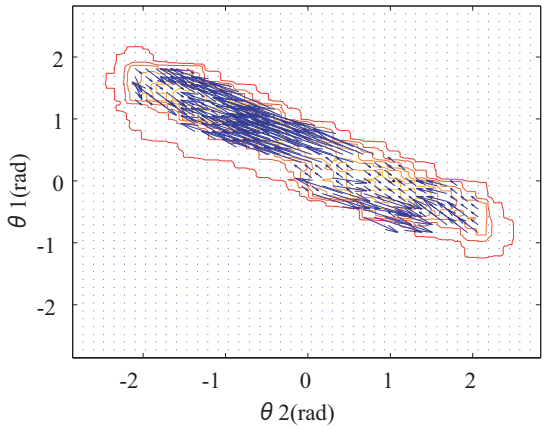

Fig. 14　200試行後の価値関数の等高線と各位置における各速度ベクトル

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．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図11,及び図12に示す．図11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．また図12から見られるように，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．図13,及び図14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大きく３つの特徴が挙げられる．

- 評価値の変動が収まり始める辺りで，目標値までの空間軌道が形成されている
- 変動が収束に向かっている最中に形成される空間軌道はほぼ一定で，軌道を通過する速度が向上
- 軌道が不安定化されたことによる次の変動の発生

人間の行動から形成された価値関数を調べることで，上記の特徴を抽出できたことから，提案した解析手法は試行錯誤の過程を調べる方法として有用であると考えられる．またこれらの特徴より，少なくとも本実験環境における被験者の学習過程において，軌道の生成と追従という段階的な手順が踏まれていることがわかった．これに加え，追従則の学習段階においては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，軌道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．

6.　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調査するため，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唆された．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．

参考文献

1) 中村仁彦:非ホロノミック系制御研究の展望，計測と制御，Vol. 36, No. 6, 384/389(1997)
2) 山田克彦:非ホロノミック系の軌道生成，計測と制御，Vol. 36, No. 6, 390/395(1997)
3) 三平満司:非ホロノミック系のフィードバック制御，計測と制御，Vol. 36, No. 6, 396/403(1997)
4) 荒井裕彦:2階の非ホロノミック系の制御，計測と制御，Vol. 36, No. 6, 404/410(1997)
5) A. Astolfi: Discontinuous control of nonholonomic systems, Systems. and Control Letters, 27, 37/45(1996)
6) K. Imafuku, Y. Yamashita, H. Nishitani: Control of a Wheeled Vehicle Using the Viscosity Solution of the Hamilton-Jacobi Partial Differential Equation, JRSJ, vol. 17, No. 5, 689/695(1999)
7) H. Inooka, Y. Shito, K. Yu: 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)
8) 谷貝将通, 石原正, 猪岡光: 非駆動関節を有する2リンクアームの手動制御, 計測自動制御学会東北支部研究集会, 206/1(2002)
9) 銅谷賢治，森本淳，鮫島和行:強化学習と最適制御，システム/制御/情報，Vol. 45, No. 4, 186/196(2001)
10) R. S. Sutton and A. G. Barto: Reinforcement Learning An Introduction, MIT Press(1998)
11) Barto, A., Sutton, R., and Anderson, C.: Neurolike adaptive elements that can solve difficult learning control problems, IEEE Trans. on Systems, Man, and Cybernetics, SMC13, 834/846(1983)
12) C. J. C. H. Watkins and P. Dayan: Q-learning, Machine Learning, Vol. 8, No. 3, 279/292(1992)
13) G. A. Rummery and M. Niranjan: On-line Q-learning using Connectionist Systems, Technical Report CUED/F-INFENG/TR 166, Engineering Department, Cambridge University(1994)
14) G. Oriolo and Y. Nakamura: Free-Joint Manipulators: Motion Control under Second-Order Nonholonomic Constraints, Proc. of IROS'91, 1248/1253(1991)