強化学習を用いたスポーツロボットの大車輪運動の獲得と その行動形態の考察

坂 井 直 樹*・川 辺 直 人**・原 正 之***
豊 田 希*・藪 田 哲 郎*

Acquisition of Robotic Giant-swing Motion Using Reinforcement Learning and Its Consideration of Motion Forms

Naoki SAKAI*, Naoto KAWABE**, Masayuki HARA***,
Nozomi TOYODA* and Tetsuro YABUTA*

This paper argues how a compact humanoid robot can acquire a giant-swing motion without any robotic models by using Q-Learning method. Generally, it is widely said that Q-Learning is not appropriated for learning dynamic motions because Markov property is not necessarily guaranteed during the dynamic task. However, we tried to solve this problem by embedding the angular velocity state into state definition and averaging Q-Learning method to reduce dynamic effects, although there remain non-Markov effects in the learning results. The result shows how the robot can acquire a giant-swing motion by using Q-Learning algorithm. The successful acquired motions are analyzed in the view point of dynamics in order to realize a functionally giant-swing motion. Finally, the result shows how this method can avoid the stagnant action loop at around the bottom of the horizontal bar during the early stage of giant-swing motion.

Key Words: reinforcement learning, giant-swing motion, non-Markov property

1. はじめに

これまでに著者らは強化学習を適用して芋虫型ロボットやヒトデ型ロボットの行動獲得を試みてきた^{1),2)}。これら過去の研究はロボットの動作を準静的に拘束し、状態遷移が決定論的に行なわせることによってマルコフ決定過程の性質が満たされる静的なタスク(マルコフ性)であった。マルコフ性を確保したことによって強化学習が適用可能となり、ロボットの前進・回転行動を実現するとともに、得られた行動形態を定量的に解析することを行なった。そこで、本論文では状態遷移が環境とロボットとの相互作用の不確定性の影響のために確率的になり、マルコフ決定過程の性質が満たされない動

的なタスク(非マルコフ性)の行動獲得へと議論を進展させるため、ロボットの大車輪運動に着目した。

多自由度開ループ制御機構のロボットによる大車輪運動は遠心力、コリオリ力、重力による非線形性をもつ。また手首に相当する第1関節が非駆動関節であり、非ホロノミック拘束となっているため、非ホロノミック拘束を有する劣駆動システムに分類される。このようなシステムの運動制御は非ホロノミック制御とも呼ばれ、さまざまな制御手法により大車輪運動を行なわせる研究は過去にいくつか報告されている^{3)~5)}。しかしながら、一般的に制御系を適用した場合、得られる行動は設計者の予想の範疇に収まり、その行動が最適であるということを保証できない。一方で、学習によって大車輪運動の獲得を試みた例もある。西村らは複数の制御器を用意し、それらの切り換えのタイミングを学習させることでアクロバットの振り上げ・安定化に成功している⁶⁾。鯨島らは、複数のモデルとコントローラの組の中から最適なものを選択することによって、単振り子の振り上げをシミュレーションで検証している⁷⁾。また大車輪運動以外においても学習によって行動の獲得を試みた例もあり、福田らはブラキエーションロボットの制御において制御器の切り替えを強化学習によって行なっている⁸⁾。しかしながら、これらの研究においても運動方程式が既知のシミュレーションであった

* 横浜国立大学大学院工学府 横浜市保土ヶ谷区常盤台 79-5

** 東京大学大学院情報理工学系研究科 東京都文京区本郷 7-3-1

*** スイス連邦工科大学ローザンヌ校 Station 9, CH-1015 Lausanne, Switzerland

* Graduate School of Engineering, Yokohama National University, 79-5 Tokiwadai, Hodogaya-ku, Yokohama

** Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo

*** Swiss Federal Institute of Technology, Station 9, CH-1015 Lausanne, Switzerland

(Received July 17, 2009)

(Revised October 23, 2009)

り、複数のコントローラがあらかじめ用意されており、内部モデルが全くない状態から学習した研究はほとんどない。また、コントローラ的设计は设计者の思想に大きく依存するため、予想外の行動の発現は期待できない。

森本らは強化学習を階層化することで、2 自由度ロボットの起き上がり運動について実ロボットを用いて成功させている^{9),10)}。森本らの実ロボットを用いた研究成果以外は、上記に示すようにロボットの動的なタスクに関しては、シミュレーションベースの研究がほとんどで、実ロボットにシミュレーション結果を適用してその問題点を議論している研究は非常に少ない。本論文では内部モデルなどのロボットに関する情報を一切用いずに、森本らの巧妙な方法とは異なり、強化学習分野の Q 学習をシンプルに実ロボットに適用した際の問題点の解明を行なったものである。実ロボットとしては、環境との相互作用のみで人間に似た関節配置と可動範囲をもつ 5 自由度の実ロボットの大車輪運動の獲得を実現している。具体的には、シミュレータを用いて強化学習を行ない、獲得した行動形態を実機にフィードバックして学習の有効性について検討を行なっている。また、学習によって獲得された行動について力学的な観点から解析を試みている。最終的に、強化学習を動的なタスクと静的なタスクに適用した場合の相違点を比較することで、動的なタスクに対する強化学習の諸問題を整理するとともに、本論文で用いた Q 学習がロボットの太車輪運動の獲得に有効であることを示している。

2. 太車輪ロボットの实验システム¹¹⁾

2.1 实验システムの構成

強化学習を用いて太車輪運動を実現するために、Fig. 1 に示すスポーツロボットを製作した。このロボットは 5 つのアクチュエータ (Best Technology 社, Dynamixel AX-12+) から構成され、それぞれ人間の腰、股、膝関節を模倣している。鉄棒と腕との間にはベアリングがはめ込まれており、フリースポイントとなっている。製作したロボットの各リンク長と質量は Table 1 に示すとおりである。Fig. 1 のロボットを含む強化学習の实验システムの全体像を Fig. 2 に示す。

Fig. 1 Giant swing robot

光点位置計測装置の 1 つである PSD センサを用いている。鉄棒の支柱とロボットの第一リンクに LED ターゲットを取り付けてロボットの位置計測を可能にしている。

2.2 学習用動力学シミュレータ

強化学習のアルゴリズムを実機に適用して学習を行なう場合、Q 値が十分に更新されるまでに約 50 万回の学習が必要となり、ロボットの疲労試験の様相を呈し、ロボットが破壊するので学習用シミュレータが必要となる。本論文ではオープンソースの 3 次元物理計算エンジン Open Dynamics

Table 1 Link parameter values

Link number	Length [m]	Weight [kg]
1	0.141	0.197
2	0.074	0.195
3	0.068	0.146
4	0.046	0.174

Fig. 2 Experimental system

Fig. 3 Robot in simulation

Fig. 4 Comparison of motor performances

Engine(ODE)を用いて、Fig. 3 に示すようにシミュレータ内に実機の大車輪ロボットを再現した。シミュレータを実機性能と一致させるためにシミュレータのアクチュエータのパラメータを実機のアクチュエータの応答と一致するように調整した。Fig. 4 に $0, \pi/6, \pi/3, \pi/2, 2\pi/3$ rad を目標角度としてランダムに実機およびシミュレータのアクチュエータに与えたときの応答を示す。両者を比較すると特性はほぼ一致していることがわかる。

3. 強化学習アルゴリズム

3.1 Q 学習アルゴリズム

強化学習^{12),13)} の1つである Q 学習は、状態と行動によって定義される行動価値関数(Q 値)を更新することによって、最終的に最適な行動を獲得するというアルゴリズムである。マルコフ性が確保されるという条件のもとでその収束性は保証される。Q 値の更新式は以下で表わされる。

$$Q(s_t, a_t) \leftarrow (1-\alpha) \cdot Q(s_t, a_t) + \alpha \cdot \Delta Q \quad (1)$$

$$\Delta Q = r_{t+1} + \gamma \max_a Q(s_{t+1}, a_t) \quad (2)$$

ここで、 s_t, a_t, r_t はそれぞれ時刻 t におけるエージェント(ロボット)の状態、行動、および環境から受け取る報酬を表わす。 α は学習率、 γ は割引率である。割引率に関しては、通常 1 に近い値で学習を行なうことが多いが、1 に近づけすぎると逆に学習の収束が遅くなることが指摘されているため¹⁴⁾ $\gamma=0.9$ とし、学習率を $\alpha=0.9$ と設定した。

ϵ -greedy 行動選択では $\epsilon(0<\epsilon<1.0)$ の値によってランダムな探索行動と最適行動の選択を行なっている。すなわち、 $\epsilon=1.0$ のときは確率 1 でランダムな探索行動を、 $\epsilon=0.0$ のときは確率 1 で最適行動を取る。この値を時間とともに変化させることによって、Fig. 5 のように探索行動から最適行動へと学習の進行に伴って行動選択の確率を変化させることができる。

3.2 Q 学習の適用

エージェント(ロボット)の状態は、ロボットの第一リンクに相当する腕部の角度と角速度を用いて定義する。Fig. 6 (a) に示すように角度を 24 等分し、第一リンクがどの領域に含まれるかを観測して角度状態を決定している。さらに、Fig. 6(b) に示すように第一リンクの角速度を 6 分割し、位置と速度の組み合わせで状態を定義している。 $\theta_{\max}$ はロボットを $\theta_0=\pi$ rad から自然落下させたときの $\theta_0=0$ rad で

の角速度を用い、 $\theta_{\max}=10$ rad/s とする。よって、 $24 \times 6=144$ の状態が定義される。

つぎに行動を定義する。ロボットは 5 自由度を有しているが両足の股、膝関節には等しい角度を指令するため、実際には 3 自由度として扱うことができる。Fig. 7 のように人間と同じような関節の可動範囲を設定し、 $\pi/6$ rad 刻みで行動を定義すると 40 パターンとなる。また、5 章で用いる行動の

(a) Definition of the position state

(b) Definition of the velocity state

Fig. 6 Definition of the robot state

Fig. 7 Movable ranges of each joint

Fig. 5 ϵ -greedy action selection

Fig. 8 Averaged performances of Q value

番号と姿勢との関係は以下の式で表わされる.

$$Action = 20 \cdot i + 4 \cdot j + k \quad (3)$$

本論文ではロボットの内部モデルを用いず, 角度や角速度という連続空間を状態分割している. また, 状態数の増加を抑えるためにQ学習をシンプルに適用しており, 状態分割に腕部の情報しか考慮していない. さらに, 行動についてはダイナミクスに直接関係するトルクを用いずに, 姿勢を用いているため, 時刻 t で状態 s_t にあるとき, 行動 a_t を取ることで遷移する時刻 $t+1$ での状態 s_{t+1} が一意に定まらない. つまりマルコフ決定過程の性質が満たされない動的なタスクの性質をこの実験システムはもっている. このためQ値は揺らぎをもち一定の値に収束しない. そこで(4)式のようにQ値の平均値を計算し, ϵ -greedy 行動選択時に用いている. (4)式において, $\overline{Q_T}$ は更新回数 T における平均後のQ値を表わしている.

$$\overline{Q_T} = \frac{Q_T - (T-1)\overline{Q_{T-1}}}{T} \quad (4)$$

Fig. 8 に示すようにQ値を平均化することにより, 動的なタスクにおけるQ値の揺らぎを抑えられていることがわかる.

4. 大車輪運動の獲得

ロボットの学習プロセスは長時間に及び, 実機を用いて学習を行なうとある種の疲労試験となる. このため, 学習を2.2節に示した動力学シミュレータで行ない, その学習結果を実機にフィードバックして学習結果の有効性について検討を行なう. 過去の研究では部分観測マルコフ決定過程下での強化学習を適用可能な方策改善の方法として, 方策勾配法が提案されている^{15),16)}. また森本らは評価関数の関数近似を行ない, 高次元連続状態空間中での起き上がり運動をロボットに獲得させることに成功している^{9),10)}. これらに対し, 本論文では状態と行動をシンプルに設定して, 報酬操作を工夫することで動的なタスクにおける行動獲得を目指す.

4.1 報酬による学習結果の差異

Fig. 9 に Fig. 1 で示したスポーツロボットを簡単にモデル化したものを示す. Fig. 9 において θ_0 が $\theta_0 = \pi$ rad となるとときに大車輪運動は実現される. ここでは, この条件を満たす方策・報酬について考える. 本Q学習において報酬として与えた物理量を Table 2 に示す. ただし, $\theta_{f,t}$ は時刻 t にお

Fig. 9 Kinematics of the giant-swing robot

けるロボットの足先の振れ角を表わす. また, E は系全体のエネルギーを表わす. 基準座標系を Σ_0 として, 慣性座標系で移動している点 A 回りの質点系の運動エネルギーは

$$K = \frac{1}{2} m_i^0 v_A^2 + m_i^0 v_A \cdot v_{C/A} + \frac{1}{2} \sum_i m_i^0 v_{i/A}^2 \quad (5)$$

と表わされる¹⁷⁾. ここで, $^0 v_A$ は点 A の速度ベクトル, m は質点系の質量, $^0 v_{C/A}$ は質点系の質量中心の速度ベクトルである.

剛体の運動エネルギーは質点の運動エネルギーの総和であるから, (5)式より大車輪ロボットの各リンクの運動エネルギー K_i は, $i=1$ において

$$K_1 = \frac{1}{2} I_1^0 \omega_1^2 \quad (6)$$

$i=2, 3, 4$ において

$$K_i = \frac{1}{2} m_i^0 v_i^2 + m_i v_i (^0 \omega_i \times r_{C/i}) + \frac{1}{2} I_i^0 \omega_i^2 \quad (7)$$

と求めることができる. つぎに位置エネルギーを表わす. 各関節を θ_0 から順に角度を変化させていくことを考えると, 位置エネルギー基準点を $\theta_0 = 0$ rad で $U=0$ とすると, U_i は $i=0, 1, 2, 3$ において

$$U_i = \left(\frac{m_{i+1}}{2} = \sum_{j=i+2}^4 m_j \right) \cdot l_{i+1} \cdot g \cdot \left(1 - \cos \left(\sum_{j=0}^i \theta_j \right) \right) \quad (8)$$

以上の(6), (7), (8)式により, 系全体のエネルギー E は

$$E = \sum_{i=1}^4 K_i + \sum_{i=0}^3 U_i \quad (9)$$

と表わすことができる. Table 2 に示した系全体のエネルギーは, (9)式に示したものをを用いている.

Table 2 に示した報酬で学習させ, シミュレータ上で適用した結果を Fig. 10 に, 実機に適用した結果を Fig. 11 に示

Table 2 Reward patterns

Absolute value of swing angle	$ \theta_0 $
Absolute value of foot angle	$ \theta_f $
Swing angular velocity	$\dot{\theta}_{0,t} = \theta_{0,t} - \theta_{0,t-1}$
Foot angular velocity	$\dot{\theta}_{f,t} = \theta_{f,t} - \theta_{f,t-1}$
Energy of all system	E

Fig. 10 Swing angles from 0 radians in the simulation

Fig. 11 Swing angles from 0 radians applying learning results to the real robot

Fig. 12 State transition

す。このように Table 2 に示すすべての報酬において、シミュレータ、実機ともにロボットは角度の小さい位置で停滞してしまうことから、この実験システムがマルコフ決定過程の性質が満たされない動的なタスクの性質をもつために大車輪運動には至らないことがわかる。しかし従来研究では、梶原らは2リンクロボットの内部モデルを用い、エネルギー変化率が大きくなるように共振現象を利用して制御することで大車輪運動を実現させている⁵⁾。このように、エネルギーを増加させるように学習をすれば大車輪運動を獲得できると考えられたが、実際にはこれとは異なる結果が得られた。これはロボットにエネルギーを増大させるように学習させても、腕部の角度を増大させるのに有効である特徴的なQ値が存在せず、Fig. 12 に示すように状態遷移が停滞ループを形成し、このループから脱却できないためである。Fig. 12 は報酬が腕角度の絶対値 $|\theta_{0,t}|$ であるときの実験結果であるが、エネルギーを用いた場合もほぼ同様な傾向が見られる。角度の小さい位置で停滞してループを形成している部分は Fig. 12 のように4つの状態のみを遷移しており、4つの行動パターンを選択のみで振幅を増大させなければならず、停滞ループの振幅がおよそ 0.25 rad であることから、Fig. 6(a) の角度の状態分割 $\pi/12 \text{ rad} (=0.26 \text{ rad})$ の初期位置近傍での状態分割の区分けが大きいためにタスク達成が不可能であることが考えられる。またサンプリング周期が 250 ms と長く、速度は位置の差分で計算しているため、速度の正負が切り換わる時に発生する不完全知覚問題などが原因として挙げられる。つぎに Fig. 12 に示した初期位置近傍での停滞ループの影響について検討するために、腕角度が $\pi/4 \sim \pi/3 \text{ rad}$ 付近の位置を初

Fig. 13 Swing angles from $\pi/4$ radians applying learning results to the real robot

Fig. 14 Bodily characteristics

Fig. 15 Effect of forced oscillation applying learning results to the real robot

期位置として運動を開始してみた。この結果を Fig. 13 に示す。このように角度の小さい位置でのループを回避すれば、報酬が $|\theta_{f,t}|, \theta_{0,t}$ の場合において大車輪運動を達成できることがわかった。

このように、さまざまな物理量を報酬として与えて大車輪運動の実現を目指したものの、初期位置が腕角度 0 rad から運動を開始すると停滞ループに入ってしまうタスク達成には至らないことが判明した。これらの報酬では動的なタスクによるQ値の揺らぎを克服できずに初期位置近傍では最適行動を獲得できない。そこで、初期位置近傍における振幅を増大させるために有効な報酬を考えていく。梶原らの結果⁵⁾ から共振現象が有効であることが示唆されるため、強制振動を学習により行動獲得をさせるために報酬として表わす。強制振動の外力を周期的加振力とすると Fig. 14 のように運動することが考えられる。Fig. 14 の両端の姿勢に着目し、ロボットの足先の角度 $\theta_{f,t}$ を腕部の角度 $\theta_{0,t}$ よりも振り上げていることから、報酬としては $\theta_{f,t} - \theta_{0,t}$ が考えられるが、 θ_0 の正負の区別をなくすために以下の報酬を用いる。

$$r_t = |\theta_{f,t}| - |\theta_{0,t}| \quad (10)$$

(10)式の報酬で学習したものを実機に適用した結果を **Fig. 15** に示す. このように大車輪運動の達成には至らないが Table 2 で示した5つの報酬と比べると, 3.2節(3)式に示した行動3と36を交互に選択することで, 腕部の振幅を $\pi/4$ rad を超える位置まで増大させることに成功していることから, 学習によって想定した行動選択が行なわれていることがわかる.

4.2 報酬の操作

前節の結果・考察から, 腕角度の小さい位置では(10)式の報酬が有効であり, 腕角度が大きい位置では $|\theta_{f,t}|, \theta_{0,t}$ といった報酬が有効であることが判明した. そこで, 初期位置が 0 rad からでも大車輪運動を可能とするために以上の報酬を1つに統合することを考える. そこで3つの報酬の中で次元が同じである, 足先角度の絶対値の報酬 $|\theta_{f,t}|$ と強制振動の報酬 $|\theta_{f,t}| - |\theta_{0,t}|$ の2つに着目する. この2つの報酬について, 腕角度が変化したときの報酬の値の傾向を **Fig. 16** に示す. ここで $|\theta_{f,t}|$ の値はロボットの取る姿勢により変化するので, 主としてロボットの初期姿勢である3.2節の(3)式に示した行動の4を用いた ($\theta_{0,t} = \theta_{f,t}$). また, 強制振動の報酬の値については, ロボットの姿勢の中で $|\theta_{f,t}| - |\theta_{0,t}|$ の値が最大である3.2節の(3)式に示した行動の36の値を用いて計算した. **Fig. 16** から読み取れることが2つあり, それはそれぞれの報酬の傾きと大小関係である.

$$\begin{cases} |\theta_{f,t}| - |\theta_{0,t}| = \text{const.} \\ |\theta_{f,t}| \propto \theta_{0,t} \end{cases} \quad (0 \leq |\theta_{0,t}| \leq \pi) \quad (11)$$

Fig. 16 Reward value by arm angle

Fig. 17 Achievement of giant-swing applying learning results to the real robot

$$\begin{cases} |\theta_{f,t}| - |\theta_{0,t}| > |\theta_{f,t}| & (|\theta_{0,t}| < \pi/4) \\ |\theta_{f,t}| - |\theta_{0,t}| < |\theta_{f,t}| & (|\theta_{0,t}| > \pi/4) \end{cases} \quad (12)$$

ここで(12)式を見ると, 前節で判明したタスク達成に有効な報酬において, 腕角度の範囲が小さいとき ($0 < \theta_0 < \pi/4$) は強制振動の報酬が有効で報酬の値も大きく, 腕角度が大きいとき ($\pi/4 < \theta_0 < \pi$) は足先角度の絶対値の報酬が有効で報酬の値も大きいことがわかる. 最適行動を決定するQ値は報酬の値によって決まるので, 強制振動の報酬と足先角度の絶対値の報酬の2つの報酬を足し合わせて1つの報酬にしても, **Fig. 16** の傾向から両方の報酬の性質を失わずに保存できると考えられる. この仮説から, 1つの報酬で $\theta_0 = 0$ rad から大車輪運動の獲得が可能になると考え, 次式に示す報酬を用いて検討を進めた.

$$r_t = 2 \cdot |\theta_{f,t}| - |\theta_{0,t}| \quad (13)$$

(13)式に示した報酬で学習したものを実機に適用した結果を **Fig. 17** に示す. 報酬を統合したことにより, 1つの報酬で初期位置が $\theta_0 = 0$ rad からの大車輪運動の獲得に成功している. このことから両方の報酬の値に差が存在すれば, このような報酬の操作をできることが示唆される.

5. 行動形態の考察

5.1 大車輪運動の解析

4.2節において学習によって獲得された大車輪運動の行動形態を解析し, 学習結果の物理的な意味を明らかにすることを試みる. 解析するデータを収集するために, 同一の学習結果を用いて1回分を $\theta_0 = 0$ から開始させて $\theta_0 = \pi$ rad となったときを終了として, このプロセスを20セット行なった. その結果を **Fig. 18** に示す. ハードウェアが破損しない限り, タスク達成が失敗に終わる試行は存在しない. 大車輪運動には前回り (FR) と後ろ回り (BR) の2通りが存在するが, どちらも10回となり, これらはほぼ等しい確率で起きることがわかる. また, 横軸は大車輪運動を達成するまでに要する時間を表すが, この時間にはばらつきが見られる. この同一の学習結果を用いても, 大車輪運動を達成する時間のばらつきの要因については, 状態分割の粗さ, 不完全知覚問題やアクチュエータが位置制御機能のみでトルク制御機能を有しないために, ダイナミクス制御機能を有せず, 動的なタスクの特徴を有する問題となり, それが確率的な揺らぎの要因となっていると考えられる.

Fig. 17の学習結果を実機に適用した結果のタスク達成ま

Fig. 18 Time of giant-swing motion

での過程を特徴的な行動形態を基準に3つの段階に分けて考察する。第1段階を運動開始から4秒程度、第2段階を約 $\theta_0 = \pm\pi/3$ radで運動する範囲、そして最後の大車輪運動に至る過程を第3段階に分割している。まず第1段階の行動形態について考察する。4.2節より、腕部の振幅が小さいときには、強制振動をするような行動を獲得するように報酬を与えている。そのようすを観察するためにロボットの姿勢の変化を表現しているFig. 9に示す角度 φ に着目した。腕部の角度と角度 φ のグラフをFig. 19に示す。 φ の波形は正弦波的な特徴を有しており、この実験結果の実際の姿勢は正弦波的な挙動に見え、ロボットは強制振動により腕部の振幅を増大させていることがわかる。

つぎに第2段階の行動形態について考察する。Fig. 17における範囲は5~25 secの間である。第2段階の行動形態をFig. 20に示す、回転方向に対し行動を(3)式で規定される37→17→36と選択することによって、腰部に相当するモータでトルクを出力している。その後の行動は36,38といった慣性モーメントの小さい姿勢を維持している。他の特徴的な姿勢として、回転方向に対し行動を36→32→36と選択する場合もあり、股部に相当するモータでトルクを出力していた。またトルクを出力する前は慣性モーメントの小さい36の行動を選択している。このように第2段階では、ロボットは慣性モーメントの小さい姿勢(行動)を維持しながら、数秒に1回、腰もしくは股関節の1つだけを $\pi/3$ rad動作させてモータの最小動作角度でトルクを出力し、振幅を少しずつ増加させていた。続いてFig. 17に示すように、25 secのときに腕部の振幅がそれ以前に比べ急に増大して第2段階から第3段階に移行していることがわかる。この状態は、第2段階のプロセスで次第に振幅が大きくなり36→37の行動選択後、 P_8 の状態に入ったときに出現するものである。Fig. 21に状態と行動を示すが、37→12→37の行動で、第2段階で示し

Fig. 19 Appearance of forced oscillation

Fig. 20 Motion forms of second stage

た1つのモータの駆動ではなく、腰と股関節の2つのモータを駆動することにより、一気に振幅を増大させている。

最後に第3段階の行動形態について考察する。第3段階は大車輪運動に至る最後の過程であるため、行動はFRとBRの2パターン存在する。また、第3段階では回りそうで回らないような状況(ループの発生)が起こり、その後大車輪運動の達成に至るが、ループの中の繰り返し運動は特徴的なものではないので第3段階として解析するのはタスク達成の最後の半周期とする。第3段階では状態と行動選択の間にある傾向が見られ、確率的に支配的な行動形態と、そうでないものとに分かれた。そのうち確率的に支配的なものについて解析する。まず、FRについては、支配的な行動形態が90%となり、その状態と行動をFig. 22に示す。このように①で股部のモータを動かすことでトルクを出力して勢いをつけ(出力後の姿勢は慣性モーメントが小さい)、③では腕の位置を上げて $\theta_0 = \pi$ radとするために伸びるような姿勢を取っている。②は③への移行段階だと考えられる。つぎにBRについては、支配的な行動形態が70%となり、その状態と行動をFig. 23に示す。このように①ではロボットの姿勢は伸びたまま、②で腰と股部を曲げることで慣性モーメントを小さくして回転しやすくし、③で股と膝部のモータを動かすことでトルクを出力して大車輪運動を達成している。以上をまとめると、FRは $\theta_0 = \pi$ radとなる半周期前にトルクを出力し、その勢いで一気にタスクを達成するが、BRは $\theta_0 = \pi$ radとなる間近でトルクを出力しており、FRとBRではトルクを

Fig. 21 Action selection for third stage

Fig. 22 Motion forms of the forward rotation

Fig. 23 Motion forms of the back rotation

出力するタイミングが異なることが判明した。これは慣性モーメントが小さいと回転運動しやすいことを意味し、ロボットの姿勢は前後非対称となっている。このため、ロボットが慣性モーメントの小さい姿勢を極力長時間維持するために、以上に示した行動形態を学習により発現したことが示唆される。

5.2 動的タスクの考察

静的なタスクにおいて強化学習を適用し行動獲得を試みた例があり、本山らは芋虫型ロボットに強化学習を適用して前進行動獲得に成功している¹⁾。この結果に示されるように、静的なタスクでは獲得された行動形態が学習によって自ラリズムを作り出したために、状態遷移がループを生成している。その状態遷移を表わしたものを Fig. 24 に示す。マルコフ性が保証されているのでループに揺らぎは存在しない。本論文で扱っている大車輪運動は動的なタスクであり、アクチュエータの発生トルクに密接に関連する。しかし、本実験システムに用いられているアクチュエータは位置制御機能しか有しておらず、行動選択は姿勢の選択となり発生トルクを一意に制御することはできない。このため、現在の状態からつぎの状態は一意に決まらず確率的な挙動となり、マルコフ性が崩れることになる。ただし、確率的な挙動の中にも、Fig. 25 に示すようにループが形成されてゆくの、その確率的な挙動ループについて考察する。

5.1 節で示した第2段階の振幅があまり増大していない領域では Fig. 25 に示すように、1本のループで示される確定的なループではなく、複数のループで示される確率的な挙動の遷移ループを示している。またこのループから抜け出す部分を Fig. 26 に示す。位置の状態が8で速度の状態が1のときに30という行動を選択しているが、そのつぎの状態は一意に決まらず、同じ行動を選択しているにもかかわらず3回目によくループを抜け出すことに成功している。これらは同じ行動を選択しても、前記に示したようにつぎの状態が一意に決まらないためであり、動的なタスクの確率的な特徴が見られる。

つぎに、第3段階における大車輪運動達成以前の状態も Fig. 27 に示す。確率的な状態遷移ループが存在し、ここでも第2段階のループと同様の特徴が見られる。またループか

Fig. 24 State transition for static task

Fig. 25 Cyclic loop at the second stage

Fig. 26 Escape the cyclic loop at the second stage

Fig. 27 Cyclic loop at the third stage

ら抜け出す部分を Fig. 28 に示す。位置の状態が21で速度の状態が3のときに34という行動を選択しているが、そのつぎの状態は一意に決まらず、同じ行動を選択しているにもかかわらず3回目によくループを抜け出すことに成功している。これらにも動的なタスクの特徴が見られる。

以上をまとめると、Fig. 29 のようなことが明らかになる。ロボットは運動を開始して強制振動をさせながら第2段階を目指す。つぎに第3段階へ移行するために通らなければならない特定の状態(位置状態: P_8)になるまで Fig. 20 のような行動を繰り返すことで、振幅を維持・増大させている。しかし特定の状態となっても、動的なタスクのため第3段階へ移

Fig. 28 Escape the cyclic loop at the third stage

Fig. 29 Process of giant-swing motion

行できるかは確率的となる。これによって第2段階のループが形成される。そして第3段階では大車輪運動を達成すべくさまざまな動作をしているが、前回りもしくは後ろ回りをするには、それぞれいくつかの特定の状態で特定の行動を取らなければならない。その条件が満たされるまでは大車輪運動を達成させることはできない。ここでも、特定の状態で特定の行動を取ったとしても、状態分割の粗さや発生トルクのばらつきのために、タスク達成できるかは確率的となる。そのため回りそうで回らないような状態が続くため、第3段階のループが形成されている。Fig. 17 に示した大車輪運動に至るまでの時間にばらつきがあるのは、第2段階と第3段階に存在するループを抜け出す時間が毎回異なるためにそのような結果が得られたものと推定される。

6. まとめ

本論文で用いたロボットの大車輪運動は、サンプリングの粗さから速度の正負が切り換わるときに起こる速度状態の不完全知覚問題、位置状態に腕部の角度のみを用いることによるロボットの姿勢の不完全知覚問題、アクチュエータが位置制御機能しか有せずトルク制御機能を有していないこと、およびベアリングの摩擦などに起因するさまざまな非マルコフ性に関連する動的なタスクの問題を、Q 値の時間平均化と報酬を操作することによりこれを克服した。運動開始からタスク達成までに、報酬としては2つのモードが存在し、始めは

強制振動をするような動き、その後足先を上げるように学習することで実現を可能とした。また、2つのモードを切り替えることなく1つの報酬で一連の行動を獲得させたので、設計者から特別なモード切り換えの情報を与えることなく実現している。すなわちロボットが自ら環境との相互作用により大車輪運動を獲得した。獲得した行動形態を見ると3つの段階が存在し、状態遷移を見ると第2,3段階で確率的な停滞ループを形成しているところがある。このループは静的なタスクでも見られるものである。しかし、この動的なタスクの場合には、現在の状態と行動からつぎの状態が予測できないために、そのループは1本ではなく確率的な挙動を示す遷移ループとなる。そのようなループから抜け出るには特定の状態になる必要があるが、特定の状態となっても動的なタスクであるためにループから抜け出せるかどうかは確率的となった。そのため、大車輪運動を達成するまでの時間は一定とはならなかった。以上に示したように報酬によっては停滞ループを形成してしまいタスク達成不可能なもの、いくつかのループが存在しながらもタスク達成可能なものが存在した。また、動的なタスク特有の確率的な性質も示すことができた。今後はさまざまな報酬に関して、タスク達成可能なものと不可能なものについて、どの動的なタスクの諸問題によりそのような違いが出てくるのかを解明し、ロボットの大車輪運動以外のマルコフ性の保証されないようなタスクにも適用していきたいと考えている。

参考文献

- 1) 本山, 山科, 原, 黄, 藪田: 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, **75**-749, 3525/3532 (2006)
- 2) 鄭, 井上, 原, 黄, 藪田: 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, **72**-723, 122/132 (2009)
- 3) 南澤, 美多, I. Pantelidis, 山北: アクロボットの振り上げ倒立制御と特異点問題, 日本ロボット学会誌, **20**-2, 85/88 (2002)
- 4) 山浦, 鈴木, 小野: 3 リンク鉄棒ロボットの的大車輪運動と制御, 日本機械学会論文集 C 編, **71**-707, 2289/2296 (2005)
- 5) 梶原, 橋本, 松田, 土谷: 鉄棒ロボットの数理解析と運動制御, 日本ロボット学会誌, **18**-4, 515/520 (2000)
- 6) 西村, 吉本, 時田, 中村, 石井: 複数制御器の切り替え学習法による実アクロボットの制御, 電子情報学会論文誌 A, **J88**-5, 646/657 (2005)
- 7) 鮫島, 片桐, 銅谷, 川人: 複数の予測モデルを用いた強化学習による非線形制御, 電子情報通信学会論文誌 D-II, **J84**-9, 2092/2106 (2001)
- 8) 福田, 長谷川: 複数のコントローラの学習方法, 日本機械学会論文集 C 編, **63**-610, 2043/2051 (1997)
- 9) 森本, 銅谷: 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, **J82**-11, 2118/2131 (1999)
- 10) J. Morimoto and K. Doya: Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning, Robotics and Autonomous Systems, **36**-1, 37/51 (2001)
- 11) M. Hara, N. Kawabe, N. Sakai, J. Huang, H. Bleuler and T. Yabuta: Consideration on Robotic Giant-swing Motion Generated by Reinforcement Learning, Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 4206/4211 (2009)

- 12) 三上, 皆川: 強化学習, 159/161, 森北出版 (2000)
- 13) 吉本, 銅谷, 石井: 強化学習の基礎理論と応用, 計測と制御, **44**-5, 313/318 (2005)
- 14) S. Mahadevan: Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results, Machine Learning, **22**, 159/196 (1996)
- 15) 木村, 山村, 小林: 部分観測マルコフ決定過程下での強化学習—確率的傾斜法による接近, 人工知能学会誌, **11**-5, 761/768 (1996)
- 16) T. Matsubara, J. Morimoto, J. Nakanishi, M. Sato and K. Doya: Learning CPG-based Biped Locomotion with a Policy Gradient Method, Robotics and Autonomous Systems, **54**, 911/920 (2006)
- 17) 有本 卓: ロボットの力学と制御, 26/35, 朝倉出版 (2002)

~~~~~

[著 者 紹 介]

坂 井 直 樹

1987年2月23日生。2009年横浜国立大学工学部生産工学科卒業。現在、横浜国立大学大学院工学府在学中。強化学習を用いたロボットの行動獲得に関する研究に従事。日本機械学会、日本ロボット学会の学生会員。

川 辺 直 人

1985年5月23日生。2008年横浜国立大学工学部生産工学科卒業。現在、東京大学大学院情報理工学系研究科在学中。ロボットの知能化の研究に従事。日本機械学会学生会員。

原 正 之 (正会員)

1980年4月8日生。2008年横浜国立大学大学院博士課程修了。2005年日本学術振興会特別研究員DC1。2008年東京大学工学研究科精密機械工学専攻特任助教。2009年よりスイス連邦工科大学(EPFL)サイエンティスト、現在に至る。力覚提示技術を用いた人間計測およびソフトウェアの研究に従事。博士(工学)。IEEE、日本機械学会、日本ロボット学会の会員。

豊 田 希

1979年4月7日生。2007年横浜国立大学大学院博士課程修了。2007年日本学術振興会PD、2008年横浜国立大学産学連携研究員、2009年横浜国立大学生産工学科研究教員。人間の運動に与える錯視の影響など人間計測の研究に従事。博士(工学)。日本機械学会、日本ロボット学会、自動車技術会、日本フルードパワーシステム学会の会員。

藪 田 哲 郎 (正会員)

1949年7月19日生。72年東京大学工学部機械工学科卒業。74年同大学大学院修士課程修了。同年電信電話公社(現NTT)電気通信研究所勤務。2000年横浜国立大学生産工学科教授。ニューロコントロール、ハイブリッドコントロール、学習、ハプティックインタフェース、人間計測などに興味をもつ。工学博士。IEEE、日本ロボット学会、日本機械学会、電子情報通信学会などの会員。

~~~~~