2G2-02

多関節多足ロボットの強化学習

Reinforcement Learning of Walking Behavior for a Multilegged and Articulated Robot

小菅智丈[*1] 佐久間淳[*1] 小林重信[*1]
Tomotake Kosuge Jun Sakuma Sigenobu Kobayasi

[*1]東京工業大学
Tokyo Institute of Technology

Reinforcement learning is a self-learning framework in which learners automatically obtain control rules from interactions with their environment. In this paper, I have investigate a reinforcement lerning of walking behavior for a multilegged and articulated robot. It was said that the more dimentional state and action space increase, the more difficulty increase. The experimental result show that if robot structure is simple, that is untrue, and the difficulty depends on adequacy of motor power for model.

1. はじめに

強化学習は, 報酬という単純な指示から, それよりもはるかに複雑な制御規則を自動的に獲得するための手法として有望である [1]. また強化学習手法の中で確率的傾斜法をはじめとする政策勾配法は確率的政策のパラメータを価値関数の勾配を用いて更新していく方法であり，少なくとも局所的最適政策への収束性が保証されている．このため，ロボット等の実問題に対する適応的な制御手法として注目され，その有用性が確認されつつある [2][3].

強化学習の実ロボットへの適用は状態行動空間が連続値であり，さらに自由度が増すにつれて状態数の指数的な増加が起こるため困難とされてきた．連続な状態行動空間の扱い方として，価値関数 (value-function) を近似する方法が多く研究されてきた [4][5]. しかし，これらの方法は状態行動空間が高次元の場合，膨大なメモリを必要とし，実行が困難であった. しかし近年，政策勾配法 (policy gradient method) に基づく接近法である Actor-Critic 手法を用いて，8 自由度の 4 足歩行ロボット [6]，蛇型ロボットなどの制御規則の獲得に成功している.

本稿が扱う多関節多足ロボットでは自由度の増加が必ずしも学習を困難にしているわけではないことを実験によって示す. 実験には，基本的な機構が同じで自由度の異なる複数の多足多関節ロボットを用いる. このことによって構造ではなく自由度数の歩行動作獲得学習への影響を調べることができると考える. さらに，ロボットの歩行動作獲得学習のおいて，モーターの出力が学習の困難さに影響を及ぼしていることを示す．また，強化学習アルゴリズムを固定した場合，タスクを達成するためにはどのような自由度設計が好ましいかは定かではない. 本研究ではタスクの達成度と自由度の設計についてその指針を示す.

2. ロボットモデル

本研究ではロボットモデルとして

- 足関節に上下左右方向のモーター (1 足 2 自由度) を持った多足ロボット）(Type1)
- 上記のモデルの胴体部分に 2 自由度 (yaw,roll) の腰を持った多足多関節ロボット (Type2)

の 2 タイプを用いる (図 1 参照)．各関節は角度制御のサーボモータを持ち，各時間ステップの各目標角度を出力することで行動を制御する．また学習目標は自分の前進方向になるべくまっすぐに前進する制御規則を獲得することである．ロボットを制御するコントローラーが学習主体のエージェントとなり，ロボット本体を含めた外界が環境となる．ロボットは各関節の現在角度を状態入力とし，その状態に対して出力する関節角度の目標値への確率分布を政策関数とする．環境との相互作用を通じてそのパラメーターを改善する.

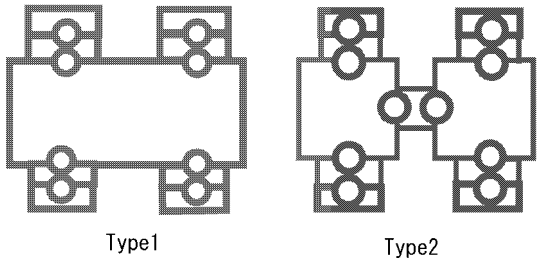

図 1: 構築した 2 タイプのロボット.Type1 を腰無し型,Type2 を腰有り型と呼ぶ.

ロボットの実装にあたっては，高精度動力学シミュレーターである Vortex を利用した (図 2 参照).

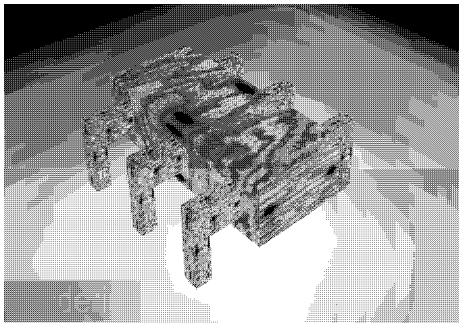

図 2: 動力学シミュレーター Vortex 上のモデル

連絡先: 小菅智丈，東京工業大学,tomotake@fe.dis.titech.ac.jp

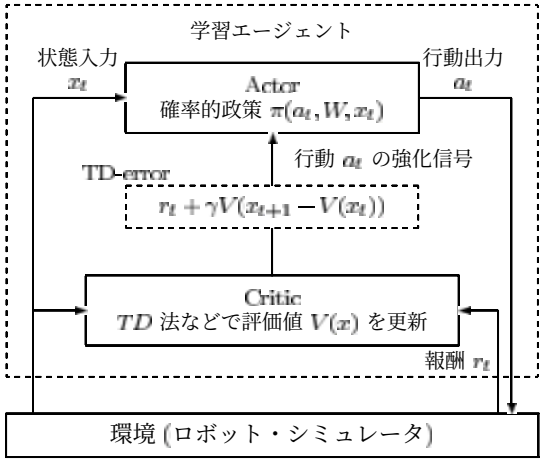

図 3: Actor-Critic アルゴリズムの一般的枠組み

3. Actor-Critic アルゴリズム

本研究では強化学習アルゴリズムとして Actor-Critic アルゴリズムを採用した．Actor-Critic アルゴリズムは状態から行動への確率分布である確率的政策を学習することができる．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できる．このことは MDP だけではなく，非マルコフ問題の一種である POMDP に対して有効であることが知られている．また，Actor-Critic アルゴリズムは，行動価値を直接学習するわけではないので，連続行動空間の学習のように行動数が無限にある場合でも対応することができる．これらの理由からロボットの学習問題に対して Actor-Critic アルゴリズムが用いられる多数の研究例が報告されている [6][7].

以下に Actor-Critic の概念図を図 3 示す．確率的政策に従って actor が行動する．その actor が選択した行動により環境が変化し，その状態が actor,critic それぞれによって観測される．その際，critic は環境から報酬を獲得する．観測された状態と獲得した報酬から式 (1) に従って TD 誤差を出力する．

$$\delta = r + \gamma V(s') - V(s) \tag{1}$$

また，内部の価値関数を更新する．actor は観測された状態と，critic から出力された TD 誤差をもとに政策を更新する．図 4 に本研究で用いた Actor-Critic の学習手順を示す.

4. 実験の目的と設定

4.1 目的と設定

本章の目的は 2 つある．1 つ目は自由度数と学習の困難さの関係を検証することと，2 つ目はモーターの出力の強度と学習の困難さの関連を検証することである．Actor-Critic アルゴリズムがどの程度の自由度まで適用できるかを調べる．Type1 については 4 足腰無し 8 自由度型,6 足腰無し型 12 自由度及び 8 足腰無し型 16 自由度ロボットを対象に実験を行う．Type2 については 4 足腰有り型 10 自由度,6 足腰有り型 16 自由度,8 足腰有り型 22 自由度ロボットを対象に歩行動作獲得実験を行う．実験に用いるロボットの仕様を表 1 にまとめて示す.

実験設定は，割引率 γ を 0.9，actor 学習率 α_π を 0.002，critic 学習率 α_v を 0.1，Actor の適正度の履歴の割引率 λ_π，Critic の適正度の履歴の割引率 λ_v はともに 1.0 とした．またロボットの腹部が床につく場合，ロボットは高さを保って前進歩行を獲得しなければならない．この場合,「高さを保つ」「前進歩行」するという様な多目的の学習問題になり問題として複雑さが増す．本研究では単目的を扱うため,「高さを保つ」タスクについては関節の可動範囲を制限することで達成することとした．よって報酬はロボットのの前進方向の距離のみで与えることとした.

1. 時刻 t において学習エージェントが環境より [-1,1] に正規化された状態 s_t を観測し，確率 $\pi(a_t|s_t)$ を用いて行動 a_t を選びこれを実行する.
2. 行動の結果得られた報酬 r_t と状態 s_{t+1} を用いて (1) 式で TD-error を計算する．ここで $\gamma(0 \leq \gamma \leq 1)$ は割引率，$\hat{V}(s)$ は割引報酬の期待値
$$TD - error = r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)$$
3. TD 法で value の推定値を更新する．ここで e_v は w の適性度 (eligibility),α_v は critic の学習率である.
$$\begin{aligned} e_v(t) &= \frac{\partial}{\partial w}\hat{V}(s_t) \\ \bar{e}_v(t) &\Leftarrow e_v(t) + \bar{e}_v(t) \\ \Delta w(t) &= (TD - error)\bar{v}(t) \\ w(t) &\Leftarrow w(t) + \alpha_v \Delta w(t) \end{aligned}$$
4. TD-error を用いて actor を更新する．ここで e_π は θ の適性度 (eligibility),α_π は actor の学習率である.
$$\begin{aligned} e_\pi(t) &= \frac{\partial}{\partial \theta} ln(\pi(a_t|\theta, s_t)) \\ \bar{e}_\pi(t) &\Leftarrow e_\pi(t) + \bar{e}_\pi(t) \\ \Delta\theta(t) &= (TD - error)\bar{\pi}(t) \\ \theta(t) &\Leftarrow \theta(t) + \alpha_\pi \Delta\theta(t) \end{aligned}$$
5. 次式で適正度をそれぞれ減らす．ここで λ_v,λ_π は crirtic と actor の適正度の割引率である.
$$\begin{aligned} \bar{e_v}(t+1) &\leftarrow \gamma\lambda_v \bar{e_v}(t) \\ \bar{e_p i}(t+1) &\leftarrow \gamma\lambda_\pi \bar{e_\pi}(t) \end{aligned}$$
6. $t \leftarrow t+1$ としてステップ 1 に戻る.

図 4: Actor-Critic の学習手順

4.2 結果

図 5,6 は各モデルの歩行動作の学習曲線である．横軸は学習ステップ数，縦軸は単位時間当たりの平均移動距離を示している．また，曲線は各モデルにつき 3 試行づつ行いその学習曲線の平均値を表している．それぞれ横軸は学習ステップ数，縦軸は単位時間当たりの平均移動距離を示している．縦軸が高ければ単位時間当たりの移動距離が高いことを示す．本実験は構造上「歩行動作」を行わなければ前進動作を獲得できない．つまり図 5,6 のグラフはそれぞれのモデルに適した歩行動作を獲得したことになる．従来強化学習は多自由度ロボットに適用が難しいとされてきた．しかしながら本実験では一般的に多自由度とされる 16,22 自由度での学習に成功している様子がわかる.

また,Type1 の四足型ロボットと Type2 の四足型のロボットの獲得された歩行動作の様子を図 7 に示す.

4.3 考察

上記のロボットモデルでは学習に失敗することはほとんどなく，ほぼ全部の試行において歩行動作を獲得した．学習曲線の形に特徴があるものの，学習に要したステップ数と自由度の数には相関がないように思われる．これらのデータは同じモーター強度を設定している．しかし、自由度設計において必ずし

表 1:
本実験で用いたロボット

vortex						
Type	1	1	1	2	2	2
自由度数	8	12	16	10	16	22
腰	無	無	無	有	有	有
足数	4	6	8	4	6	8

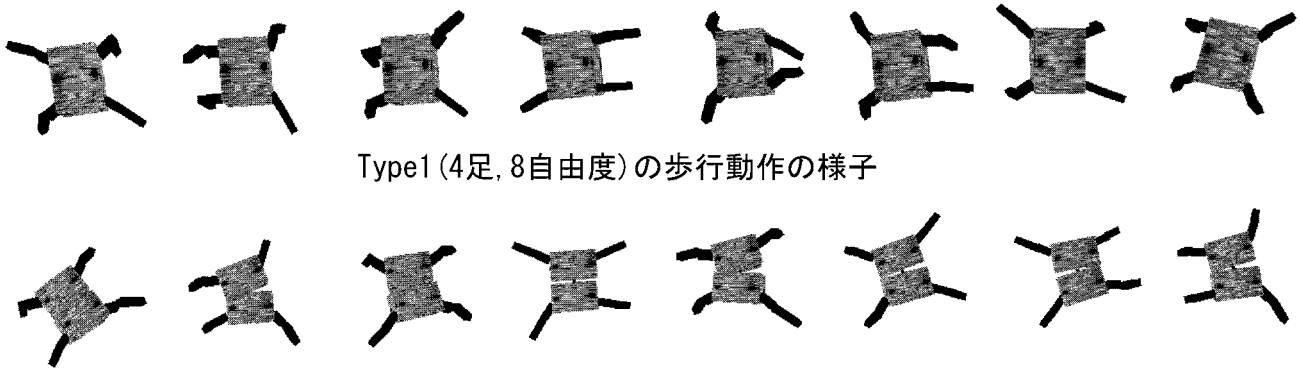

図 7: Type1 の四足型ロボット (上) と Type2 の四足型のロボット (下) の獲得された歩行動作の様子

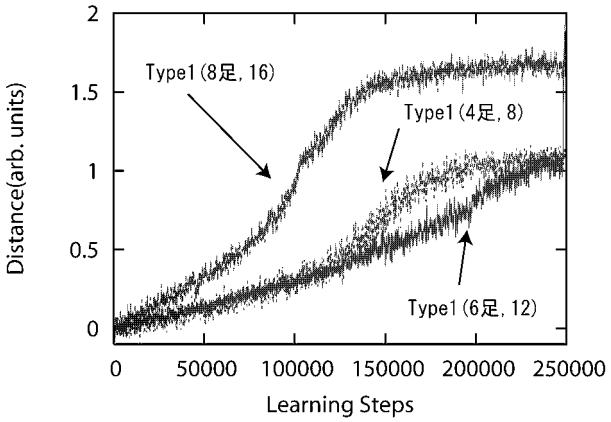

図 5: Type1 モデル学習曲線

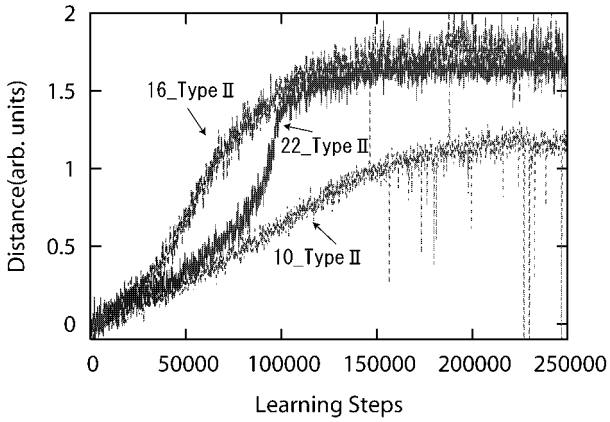

図 6: Type2 モデル学習曲線

も同じモーター強度が適しているとは限らない．モデルに対して適切なモーター強度を設定しなければ目標角度を出力しても達成できない等不都合が生じ，学習困難になる可能性もある．図 8 に Type1 の 12 自由度型ロボットにおけるモーター強度を 3 段階変化させた時の学習曲線を示す．横軸は学習ステップ数であり，縦軸は単位時間当たりの平均移動距離を示している．

図 8 から学習速度はモーター強度に関係していることがわかる．これは関節の目標角度を達成できるかどうかに関連すると考えられる．出力した関節の目標角度を達成しない場合，推定価値関数は間違った値を用いて更新されてしまう．価値関数の推定を正しく行われるためには各関節に対し出力した目標角度が確実に実行される必要がある．現在用いている状態観

測は 1 関節 2 状態であるため，目標角度と同じ状態に遷移していれば価値関数は正しく更新されると考えられる．図 8 の場合，モーター強度が弱くなるにつれ目標角度の実行率が低くなるためため学習が遅くなると考えられる．

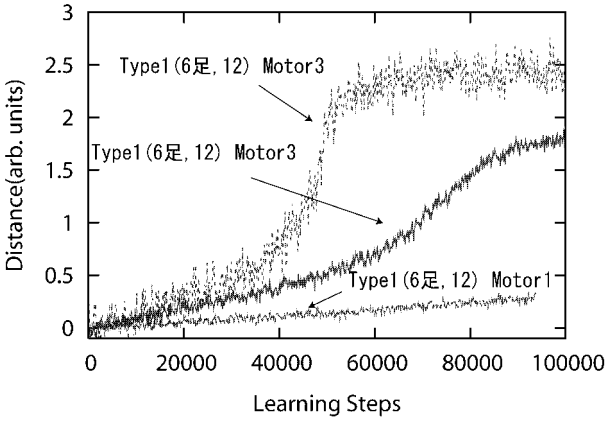

図 8: モーター強度と学習曲線

5. モーター強度とタスク達成度の関係

モデルとモーター強度の相関について実験考察をする．ロボットを制御する際には，より少ないエネルギーでより高いタスク達成度が得られることが望ましいと思われる．今回，そのような要求を満たす有効自由度の有用性を実験によって検証を行った．このような自由度設計は実機を用いることを考えた場合重要な問題となる．本実験では，Type2 の特徴である「腰」がタスクである前進歩行動作にどの程度寄与しているかを調べる．一般にモデルの自由度には独立したモーターが装備される．本実験では各自由度のモーターには同じモーターを使用し，それぞれのモーターを制御するのに必要なエネルギーは同程度であると仮定する．この場合，歩行動作に用いるエネルギーはモーター，つまり自由度の数に比例することになる．この観点から以下のような比較実験を行った．

5.1 実験設定

先での実験でのモデルのうち，Type1 の 12,16 自由度型ロボット，type2 の 10,16 自由度型ロボットについてそれぞれモーターパワーを 3 段階変化させ，歩行を学習した後の前進歩行能力について検証した．Actor-Critic アルゴリズムに用いる設定に関しては前実験と同じ設定を用いた．

6. 結果・考察

図 9 に結果を示す．横軸がモーター強度，縦軸が単位時間当たりの平均移動距離であり歩行能力を示す．図 9 の直線は各プロットの線形近似である．最高の歩行能力を得られたのは Type1 の 16 自由度型のモーター強度が 0.6 の時であるが，少ないモーター強度に対して高い歩行能力が得られたのは Type2 の方である．また、Type1 よりも Type2 の方がモーターの強度に関して依存度が強い結果を示した．Type1 は互いの自由度が独立した構造となっている．つまり互いに大きく影響を及ぼさない．Type2 は足の位置は腰の自由度と足のそれぞれの自由度の角度の組み合わせに大きく依存する．このことが，少ない自由度で高い歩行能力を実現していると考えられる．

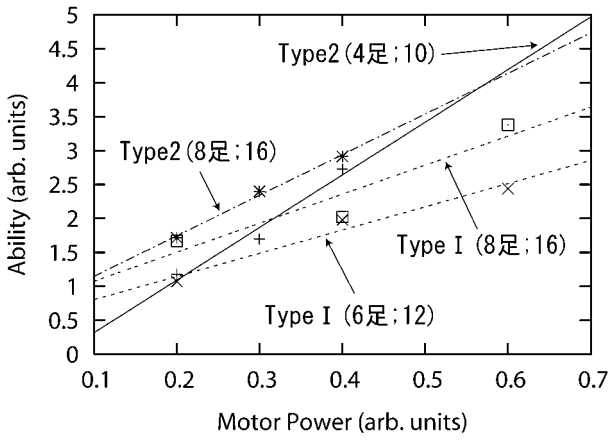

図 9: モーター強度と歩行能力の関係

6.1 おわりに

本論文では Actor-Critic アルゴリズムを用い，学習可能な自由度数とロボットモデルの自由度設計のための指針を示した．計測可能な 22 自由度まで Actor-Critic アルゴリズムが有効に働くことがわかった．また，学習の困難さと自由度数が必ずしも相関があるとは言えず，モデルに装備しているモーターの出力とモーターの相性に依存していることを実験によって示した．また，モデルによる歩行能力のモーターの出力に対する依存度を調べた結果，腰有り型ロボットが腰無しロボットよりも高いことを示した．このことが哺乳類を始めとする歩行生物に腰がついている一つの要因の可能性もある．

参考文献

[1] 木村 元，宮崎 和光，小林 重信: 強化学習システムの設計指針，計測と制御，Vol.38, No10, pp.618-623(1999).

[2] Doya, K.: Efficient Nonlinear Control with Actor-Tutor Architecture. Advances in Neural Infomation Processing Systems 9, pp. 1012-1018(1997).

[3] Morimoto, J. and Doya, K.: Acquisition of Stand-up Behavior by Real Robot using Hierarchical Reinforcement Lerning, Proceedings of the 17th International Conference on Machine Learning, pp.623-630(2000).

[4] Tsitsilis, J. N., and Roy,B.V.: An Analysis of Temporal Difference Learning with Function Approximation, IEEE Transactions on Automatic Control, Vol.42, No.5,pp.674-690(1997).

[5] Sutton, R. S and Barto, A.: Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press(1998).

[6] 木村 元，山下 透，小林 重信: 強化学習による 4 足ロボットの歩行動作獲得，電気学会 電子情報システム部門誌，Vol.122-C, No.3, pp.330–337 (2002)

[7] 角田 英太郎，青木 圭，佐久間 淳，小林 重信: 強化学習によるカササギの歩容獲得，計測自動制御学会 第 17 回自律分散システムシンポジウム，pp.159–164 (2005)