強化学習における身体成長効果の単純起き上がり問題での検証

大分大学　　○清祐大樹　杉坂政典　柴田克成

Verification of Body Growth Effect on Reinforcement Learning in a Simple Standing-up Task

Daiki KIYOSUKE, Masanori SUGISAKA and Katsunari SHIBATA, Oita University

Abstract : The authors believe that the body growth accelerates the learning of actions, such as standing up in higher forms of lives. While, the slow learning is a serious problem in Reinforcement Learning. In this paper, it was verified whether the body growth accelerates the learning in a simple standing-up problem. It was shown that even though the pendulum could not learn to stand-up when it's arm was long, it could learn when the arm became longer from short state during the learning.

１．　まえがき

人間は心身ともに非常に未熟な状態で生まれ，徐々に成長していく．なぜ，成長するのであろうか．その１つの解として，筆者らは“成長が学習に及ぼす効果”というものがあるのではないかと考えている．人間は１歳前後で歩き始めるが，この時，個人差はあるものの，身長は大人の半分以下，体重は1/5以下である．もし，成長せずに始めから大人の身体で，一から試行錯誤に基づいて，歩行の学習をしようとしても，立つことを学習するのは容易ではないと想像できる．また，学習初期の段階では失敗して転ぶことが多い．大人が立った状態から転んだ時の身体的なダメージは，子供の時とは格段に違うと考えられる．つまり，成長しながら学習することで，学習を高速化し，かつ身体に対する危険の回避にもつながっていると考えることができる．

近年，生物の行動学習からヒントを得た自律的な学習方法として強化学習が注目されている．この学習の特徴として，自らが試行錯誤を行い，自律的かつ柔軟な行動学習ができることがあげられる．しかし，その学習は探索の要素を含むため，時間がかかるという大きな問題点がある．これに対し，学習させるタスクの環境を変化させて，難易度を徐々にあげていくことが提案されている[1]．ここでは自転車の補助輪の位置を徐々に上にあげていくことで難易度を変化させながら学習させており，その結果，バランスを取りながら自転車をこぐという行動の学習が高速化されることが示されている．しかし，身体成長の効果については述べられていない．

一方，Barto らによって，倒立振子を強化学習させる際に，振り子の長さが短い方が，長い方よりもバランスをとりながら立っていることが難しいという指摘がされている[2]．前述の歩行も倒立振子を拡張したものと考えることができる．したがって，この指摘に基づくと，子供より大人の方が，学習が簡単であるということになり，筆者らの仮説と矛盾することになる．

本論文では，強化学習に身体成長を導入し，学習にどのような効果があるのかを，単純な起き上がり問題を用いて，シミュレーションで検証する．そしてさらに，Barto らの指摘が身体成長の効果を否定することにはならないことを考察に基づいて示す．

２．　強化学習

本論文では自律的な学習方法の１つである強化学習の中で，Actor-Critic アーキテクチャを用い，Critic(状態評価値)の学習には，TD（Temporal Difference）型の学習を用いて学習させた．

現在の状態評価値 P_t をもとに，TD 誤差 $\hat{r}_t$

$$\hat{r}_t = r_t + \gamma P_t - P_{t-1} \qquad (1)$$

を減少させるように，１単位時間前の状態評価値 P_{t-1} を次式によって更新していく．

$$P_{t-1} = P_{t-1} + \alpha \hat{r}_t \qquad (2)$$

ここで，r_t:報酬，γ:割引率，P_t:状態評価値，α:評価の学習係数である．また，その時の状態での Actor の出力 $\mathbf{a}_t$ に試行錯誤のための微小乱数 $\mathbf{rnd}_t$ を加えたものを実際の行動とし，(1)式で求めた TD 誤差 $\hat{r}_t$ を用い，次式によって現在の１単位時間前の行動を更新していく．

$$\mathbf{a}_{t-1} = \mathbf{a}_{t-1} + \beta \mathbf{rnd}_{t-1} \hat{r}_t \qquad (3)$$

ここで，β:行動の学習係数である．これにより P_t の時間変化が大きいと $\mathbf{a}_{t-1}$ が $\mathbf{rnd}_{t-1}$ の方向に強化され，状態評価値がより大きくなるように学習される．

３．　タスク

本論文では Fig.1 に示す単純な１リンクの振り子の起き上がりを学習タスクとして用いた．振り子が真横に寝た状態からスタートし，垂直に立つことを学習させる．状態空間は，角速度を 0.28(rad/sec) ずつ 101 分割，角度を 4 度ずつ 45 分割するため，45*101 個の状態よりなる．そして，それぞれの状態での状態評価値 P と出力トルク T をテーブルで保持し，強化学習で学習する．

身体の成長は，振り子の長さ l を 0.3m～1.0m へと，時間が 40 万秒経過した時点で最大値になる

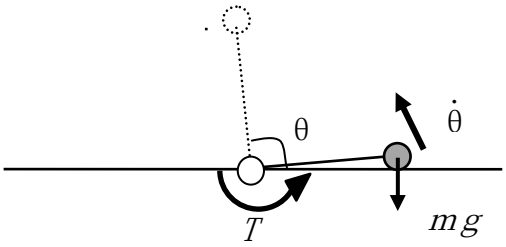

Fig.1A simple standing-up task

ように, 直線的に変化させた. そして次式に従って, その時の長さに対しての質量, 出力トルクの最大値が決まるようにした. 質量は, 密度一定で, その時の体積に比例するとして,

$$m = l^3 \qquad (4)$$

と更新した. また, 出力トルクの最大値は, 振り子を真横に浮かせた状態で静止させるためのトルクに比例するとし,

$$Tmax = T\_MAX \times ml = T\_MAX \times l^4 \qquad (5)$$

とした. ここで, m : 振り子の質量, l : 振り子の長さ, $Tmax$: 出力トルクの最大値, T_MAX: 最大値の決定係数である. ここでは決定係数を 40 とし, 振り子の長さを 0.3mから 1.0mまで成長させているので, 出力トルクの範囲は-0.324～0.324(N･m)から-40～40(N･m)と変化する. そして, 1サイクル1万秒とし, 50 サイクル, つまり, トータル時間が 50 万秒経過するまで試行を続け, 学習を行った. 1サイクル時間が経過するか, ゴール地点に到達した時に初期状態$(\theta, \dot{\theta})$=(0,0)に戻すようにし, それ以外は学習を連続して行わせた. その間, 地面に衝突した場合には速度を 0 として試行を続けた. 報酬は, 振り子が垂直から 8 度以内で, 角速度が -0.84～0.84 (rad/sec)までの状態になったとき1を与えた. また, 実際にトルクとして与える行動は

$$T = (a_{x,t} + rnd_t^3)Tmax \qquad (6)$$

としている. rnd は-1～1の一様乱数である. ここで、$a_{x,t} + rnd_t^3$ が-1 より小さい場合または 1 より大きい場合はそれぞれ-1, 1 とした. 実際に学習するのはその時点での最大トルクの何割を出力するかを決定することとなる. また, $a_{x,t}$ は-1～1 とし, 学習時にそれを超えた場合は-1 または 1 とした.

4. シミュレーション

まず, 比較のため, 振り子の長さをそれぞれ最小値 0.3m, 最大値 1.0mに固定して学習を行わせた. 次に, 振り子の長さを成長させながら学習させた. ここで, 振り子の長さ, 質量, トルクの最大値以外は変化させずに同じ条件とした. 評価の学習係数α=0.3, 行動の学習係数β=4.0, 割引率γ=0.99, 摩擦なしとした.

Fig.2 では, 成功するまでの平均ステップ数を 0.3mに固定した場合と成長した場合について示す. 1.0mの場合, 偶然何度か立つことはあったが, 学習できなかった. 成長する場合の方が振り子長を 0.3mで固定した場合より多少遅れているものの, 成功までの平均ステップ数がそれぞれ一定の値に収束していることがわかる. また, 収束した値は振り子長が 0.3mの時の方が小さい.

Fig.3,4,6,7 は, 評価と最大トルクに対する出力トルクの割合の分布を, 0.3mに固定した場合と成長した場合について示す. 図の中心部に四角で示している部分がゴールとなる. Fig.3, 6 をみると, 評価の勾配がスタート地点からゴール地点に向けて形成されていることがわかる. また, その際, 角速度はいったん上昇している. そして, それに伴って, Fig.4,7 のように行動の学習ができていることがわかる. また, 学習後の振り子の軌跡を Fig.3,6 に黒い四角でプロットした.

Fig.3,6 を比べると, 成長させた方の評価値の高い部分の角速度方向への膨らみが小さくなっている. これは, 成長するにつれて, 角速度の最大値が下っていくためである. 振り子長を固定した場合の時には, そのままでの理想的な行動を学習していくので, 評価値の高い部分の膨らみが角速度方向に, 学習が進むにつれて, 大きくなっていく. しかし, 成長しながらの場合, その時点での長さの理想的な軌道の学習が進む前に, 徐々に最大の角速度も小さくなっていく. そのため, 角速度方向に膨らんでいかなかったと考えられる. また, Fig.3,6 の右上, 左下の評価値が高くなっているのが観察できるが、振り子が起き上がってから, 地面に衝突した際に速度を 0 としたため, (0,0)または(0,π)での評価値が伝播したためと考えられる.

Fig.4,7 では正と負のトルクの分布が隣り合うようになっていて, ある程度角速度が上がったところでは, 加速ではなく減速する行動を学習している. これは, 角速度が上がり過ぎてしまうと, 最大トルクを出力しつづけても, ゴール地点までにゴールと

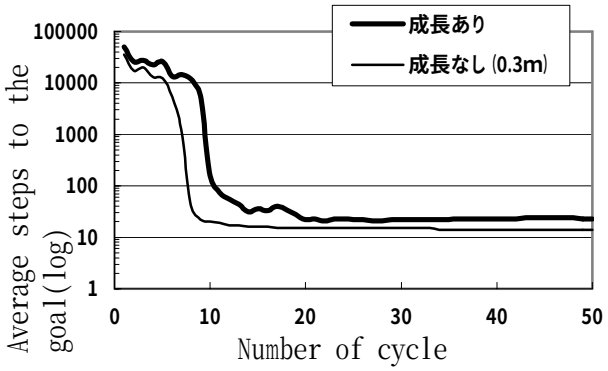

Fig.2 Comparison of the learning curve

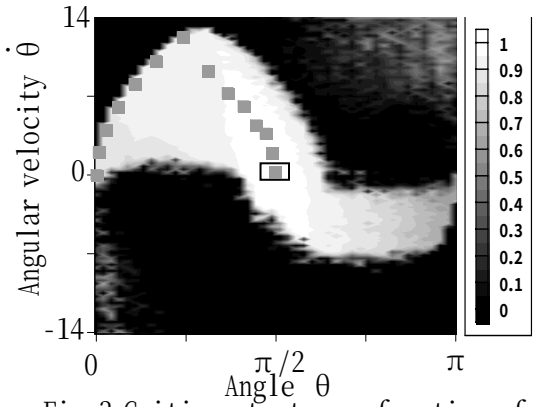

Fig.3 Critic output as a function of the pendulum state (l=0.3m)

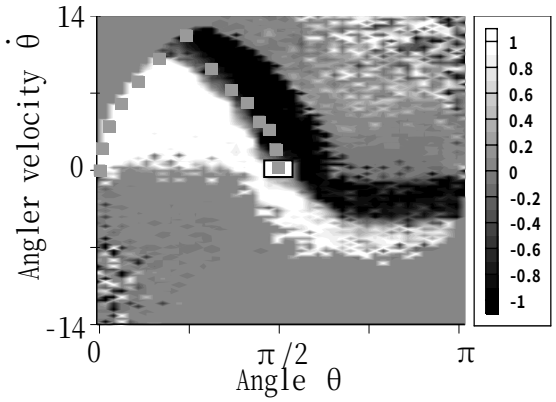

Fig.4 Actor output (action) as a function of the pendulum state (l=0.3m)

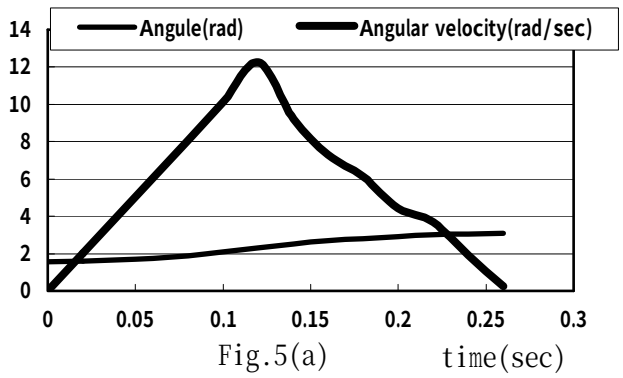

Fig.5(a)

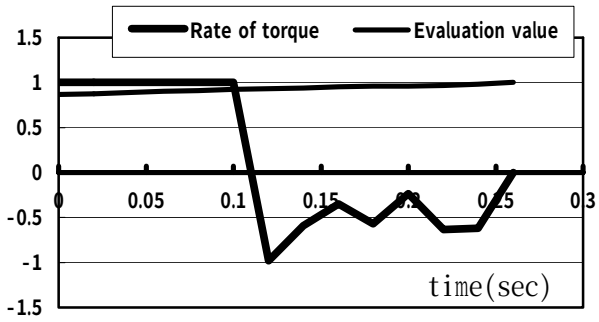

Fig.5(b)

Fig.5 Change of the variables in one trial after learning (l=0.3m)

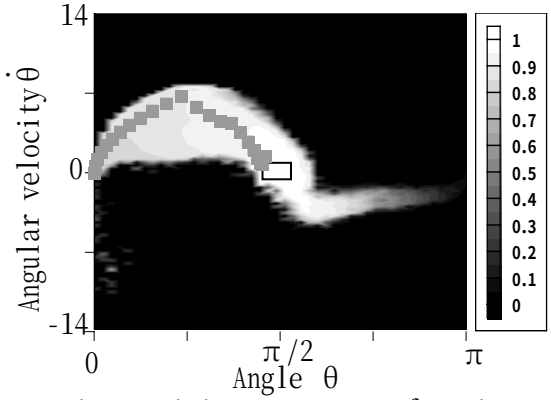

Fig.6 Critic output as a function of the pendulum state (l:0.3m→1.0m)

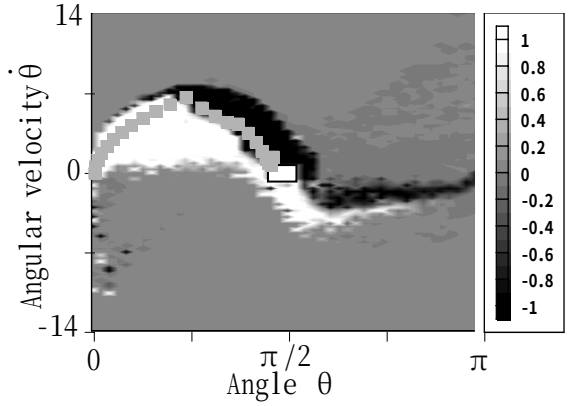

Fig.7 Actor output (action) as a function of the pendulum state (l:0.3m→1.0m)

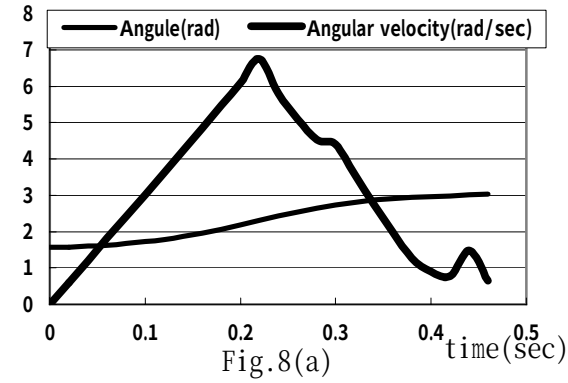

Fig.8(a)

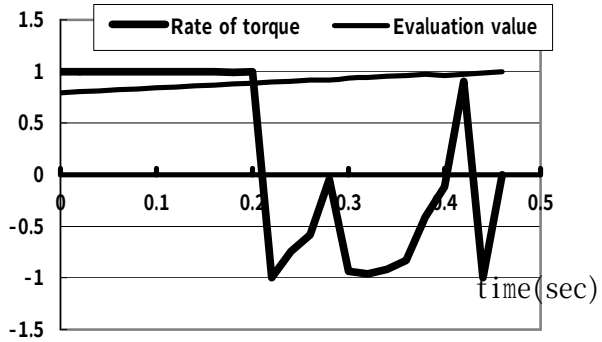

Fig.8(b)

Fig.8 Change of the variables in one trial after learning (l:0.3m→1.0m)

判定される角速度まで減速することができないためである．1.0mの場合には，立つ回数が少なすぎて，ゴールのごく近傍だけでしか評価値が高くならず，0.3mに固定した場合のように評価値の勾配がうまく得られなかったため，行動の学習も進まなかったと考えられる．

Fig.5,8 は，学習後のゴール到達までの角度と角速度の変化(a)，その時の最大トルクに対する出力トルク割合と評価値の変化(b)を 0.3mに固定した場合と成長した場合について示したものである．この図からも，いずれの場合もゴール時にちょうどよく減速するように学習ができていることがわかる．また，短いほうが角速度の最大値が大きく，到達時間も早い．

これらの結果より，振り子の長さを 1.0(m)に固定した場合では学習できなかったにも関わらず，学習することのできた 0.3(m)から，1.0(m)まで成長させながら学習することによって，1.0 の場合でも学習できるようになったことがわかる．これは，振り子が短い時での報酬が生き，評価値の差がつくため学習することができたと考えられる．また，おそらく，長いまま学習させた時でも，いずれは学習が進むようになるであろうと考えられる．したがって，成長させながら学習することで，学習の高速化を行っていると考えることができる．

5. 考察

本シミュレーションにおいては，パラメータの値を変化させると結果が大きく変わってくる．特に(5)式を用いたトルクの最大値の決定方法は本論文の結論を大きく左右する．(5)式では，前述のように，重力に対して支える力の大きさと比例するとの設定で，l^4 に比例するとした．この場合，最大の加速度の比は 1：l となるため，重力の影響を除いた時の振り子が立つまでの時間の最小値は $\sqrt{l}$ ：1 となる．これはちょうど固有周期の比と一致する．この問題では，l＝0.3 であるため，周期は 1.83 倍となるが，Fig.5,8 より学習後の到達までの時間の比もちょうどこの値に近くなっている．しかし，もし成長時に最大の角速度を揃えるように，最大トルクを調節すると，l^5 に比例することになる．この場合，重力の影響を除くと，振り子が立つまでの最短時間は等しくなる．また，重力の影響を考えると，逆に長い方が速く立つことができるようになる．

次に振り子が立つまでの最短時間と学習の関係を考える．振り子が立つまでの時間が長いということは，それだけ同一方向にトルクを出力し続けなければならない．強化学習では，乱数を用いて試行錯誤を行うため，立つまでの時間が長いほどその確率が小さくなる．この時，乱数の切り替わりの時間が長ければあまり問題にならないが，短いと少しの時間の差でゴール到達の確率が大幅に変化する．本論文のシミュレーションでも乱数の切り替わりの時間を倍にすると 1.0mで固定した場合でも学習できた．しかし，俊敏な動作を学習するためには，乱数の切り替わりの時間を短くする必要がある．

ここで l^4 の妥当性を考える．人間は筋肉によってトルクを発生するが，単純に筋肉の断面積が l^2 倍となり，筋張力が l^2 倍になると考える．さらに，モーメントアームが l 倍と考えると，トルクは l^3 倍となる．この値は関節トルクを実際に測定しなければわからないが，上記により推測すると，l^5 というのはスケーリング則から大きくはずれ，l^4 またはそれ以下の方が妥当であると考えられる．

最後に Barto らの指摘について考察する．彼らの実験では，振り子が立った状態からバランスを取って立ち続けることを目的としている．したがって，逆に今度は，振り子の長い方が，落ちるまでの時間が長くなり，その分細かい制御が可能となる．したがって，立つまでの学習とは逆に，振り子の長い方が，学習が容易となる．しかし，バランスを取る前に起き上がることを学習しなければならないが、この際には Barto の指摘は当てはまらない．さらに，Barto らのシミュレーションでは長さを変えても，力の大きさを一定としているという点でも，成長を考えた時に，不自然である．

6. あとがき

本論文で，身体成長が強化学習に対して，学習が困難であったことを可能にし，また，学習を高速化する効果があることを示した．今回の場合，成長のさせ方は，筆者らが任意に決定したものであるが，成長のさせ方により，よりよい効果が得られるかもしれない．今後は，このことも考慮した上で，より自由度が大きいタスクに応用していく．

謝辞 本研究の一部は，文部科学省科学研究費若手研究(B)(課題番号 13780295)の補助の下で行われた．ここに謝意を表する．

参考文献

[1] Jette Randlov, (2000) Shaping in Reinfo -rcement Learning by Changing the Physics of the Problem Proc. of ICML, p.p.775-782

[2] A.G.Barto, et al.(1985) TRAINING AND TRACKING IN ROBOTICS, In Proc. of IJCAI, p.p.670-672