倒立二輪型移動ロボットの
全身ダイナミクスを生かした投球動作の学習

Learning Dynamic Throwing Motion of A Wheeled Inverted Pendulum utilizing Whole Body Dynamics

○　西川　剛広　(阪大)　　高橋　泰岳　(阪大)　　中村　恭之　(和歌山大)
正　浅田　稔　(JST ERATO，阪大)　　石黒　浩　(JST ERATO，阪大)

Takehiro Nishikawa, Osaka University, 2-1 Yamadaoka, Suita, Osaka
Yasutake TAKAHASHI, Osaka University.
Takayuki NAKAMURA, Wakayama University.
Minoru ASADA, JST ERATO, Osaka University.
Hiroshi ISHIGURO, JST ERATO, Osaka University

We apply reinforcement learning to a wheeled inverted pendulum robot that acquires dynamic throwing motion utilizing whole body dynamics. Large number of parameters are needed to be calibrated so that the robot becomes able to throw a ball far away utilizing its own body dynamics while it keeps standing. We investigated the learning process of the throwing motion by application of a policy gradient method with a dynamics simulater.

Key Words: Reinforcement Learning,Wheeled Inverted Pendulum,Body Dynamics

1　はじめに

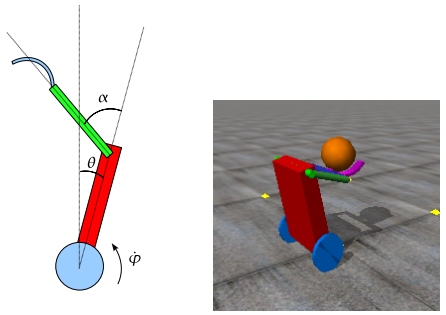
自分の身体ダイナミクスを巧みに使って様々な環境で活動できる環境適応性を持つロボットが求められている．しかし，設計の段階で環境との相互作用を完全に考慮することは困難である．そのためロボットによる行動学習が盛んに研究されている．身体ダイナミクスと環境の相互作用を有効利用する行動の一例として投球動作がある．そこで本研究では移動ロボットに投球動作を強化学習によって獲得させ，その挙動を調べた．

従来研究として，妹尾らの投球動作を行うロボットハンド [1, 2, 3] の研究がある．これらの研究では人間の投球動作時における各関節の動作を適切に連動させることで効率のよい投球動作を実現している．しかし地面に固定されているため，移動ロボットに必要な自立安定性を考慮していない．ヒューマノイドロボットはサーボ系で制御されているものが多く，身体ダイナミクスを十分に有効利用するのは難しい．また制御自由度が多いため，投球のような複雑な動作の学習は困難である．そこでよりシンプルに身体ダイナミクスを利用できると考えられる倒立振子型移動ロボットを研究対象に用いる．

2　提案手法

2.1　ロボットとタスク

シミュレーションに使用した平行二輪型移動ロボットのモデルとシミュレーターでの概観を **Fig.1** に示す．車輪と腕に自由度が１つずつあり，ボールを保持する手の部分は固定である．　ロボットはボールを持って倒立している状態から始めるとし，その状態から手を振って転倒せずに投球を行うパラメータの初期値は手動で与えた．学習は方策勾配法 [4) 5) 6)] を用いる．方策勾配法は現在のパラメータをより多くの評価値が得られると考えられる方向にパラメータを修正していくことで局所最適解を求める学習法である．ここでは Kohl and Stone [6)] によって定式化された手法を採用した．



(a) モデル　　(b) 概観

Fig.1 倒立二輪型投球ロボット

現在の方策を表すパラメータ値 $\mathbf{\Theta}$ で表す．学習には，まず現在の方策 $\mathbf{\Theta}$ を小さく変動させた T 通りの方策 $\mathbf{R}^i(1 \leq i \leq T)$ を用意する．$\mathbf{R}^i$ は $\mathbf{\Theta}$ の各成分 θ_j にランダムに ϵ_j，0，$-\epsilon_j$ のいずれかを加えて生成する．変動ステップサイズ ϵ_j はパラメータ毎に異なる値でよい．

$$R^i_j = \Theta_j + r\epsilon_j \qquad (r \text{ は-1,0,1 のいずれか}) \qquad (1)$$

次にそれぞれの方策 $\mathbf{R}^i$ にしたがって投球動作をそれぞれ 1 回行い評価を得る．用意した方策全てについて投球動作を行った後，評価関数の $\mathbf{\Theta}$ に対する勾配 $\mathbf{A}$ を近似的に求める．各パラメータ θ_j について，

- ϵ_j を加えたときの平均評価 $Avg_{+\epsilon,j}$
- 0 を加えたときの平均評価 $Avg_{+0,j}$
- $-\epsilon_j$ を加えたときの平均評価 $Avg_{-\epsilon,j}$

を求める．0 を加えたときの平均評価が最も大きい場合は，θ_j についての勾配は 0 とする．そうでない場合には，$Avg_{+\epsilon,j}$ と $Avg_{-\epsilon,j}$ の差を勾配とする．

$$A_j = \begin{cases} 0 & Avg_{+0,j} > Avg_{+\epsilon,j} \text{ かつ} \\ & Avg_{+0,j} > Avg_{-\epsilon,j} \\ Avg_{+\epsilon,j} - Avg_{-\epsilon,j} & \text{それ以外} \end{cases} \tag{2}$$

A を求めたあと，**A** を正規化し η を掛けたものに，各成分に ϵ_j の重みをつけ，**Θ** を更新する．

$$A = \frac{A}{|A|} * \eta \tag{3}$$

$$\mathbf{\Theta}_j = \mathbf{\Theta}_j + A_j \tag{4}$$

この T 回の投球動作とパラメータの更新を学習の 1 ステップとする．これを繰り返すことで，パラメータは評価が極大となる局所最適な値に近づく．

2.2 投球動作の分割 (モーションの定義)

投球動作は歩行のような反復動作ではないため，投球動作をいくつかの段階（以降モーションと呼ぶ）に分けて，そのモーション毎にパラメータを変化させることにした．本研究では投球動作を

- 投球準備のため胴体を傾けるモーション：THROW1
- 腕を振りボールを投げ始めるモーション：THROW2
- 投球中の姿勢を転倒しないよう制御するモーション：STABILIZE1
- 投球後不安定になった姿勢を制御するモーション：STABILIZE2
- 倒立モーション：STAND

の 5 つのモーションと設定した．各モーションの動きの概略図を **Fig.2** に示す．

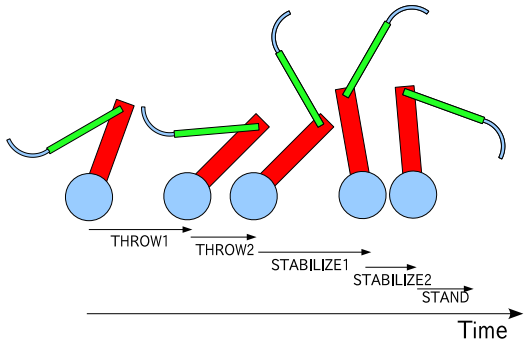


Fig.2 モーション概略

2.3 軌道生成式

各モーションにおいて胴体角，車輪角速度，腕角の目標軌道を指定する．ここでは指数関数型の目標軌道を生成することにし，現在値と最終目標値から 1 ステップ毎に次の 1 ステップでの各目標値を以下の式を用いて計算した．

$$\theta_d = (1 - w_\theta)\theta_{last} + w_\theta\theta \tag{5}$$

$$\dot{\varphi}_d = (1 - w_{\dot{\varphi}})\dot{\varphi}_{last} + w_{\dot{\varphi}}\dot{\varphi} \tag{6}$$

$$\alpha_d = (1 - w_\alpha)\alpha_{last} + w_\alpha\alpha \tag{7}$$

この w_n を以降軌道重み係数と呼ぶことにする．

2.4 学習パラメータの設定

1 つのモーションを記述するためには，モーションの継続時間，軌道生成式を計算するための胴体角，車輪角速度，腕角の軌道生成の最終目標値 θ_{last}，$\dot{\varphi}_{last}$，α_{last} と軌道重み係数 w_θ，$w_{\dot{\varphi}}$，w_α，車輪トルク制御式を計算するために胴体角，胴体角速度，車輪角速度，車輪角速度誤差それぞれのゲインである k_1，k_2，k_3，k_4，腕の P 制御の比例ゲイン k を与えればよい．このうち腕の P 制御の比例ゲイン k は固定し，それ以外の 11 個を学習パラメータとすることにした．モーションを 5 つ定義したので総計 55 個，ただし STAND モーションの最終目標車輪角速度については 0 で固定できるので，結局 54 個のパラメータについて学習を進めることにした．このうち，時間は正の値に，目標腕角 α_d は腕が車体にぶつからない範囲の，$-150° < \alpha_d < 150°$ に，各軌道重み係数 w_n は，$0 < w_n < 1$ に学習の範囲を制限してある．

2.5 トルクの制御

平行二輪型移動ロボットの車輪トルクを与える制御式は [7) 8)] に詳しい．ロボットの車輪トルクを与える制御式を以下に示す．

$$T = k_1(\theta_d - \theta) + k_2\dot{\theta} + k_3\dot{\varphi} + k_4\sum(\dot{\varphi} - \dot{\varphi}_d) \tag{8}$$

T は車輪トルク，θ は胴体角，θ_d は目標胴体角，$\dot{\varphi}$ は車輪角速度，$\dot{\varphi}_d$ は目標車輪角速度，k_n は 1 から順に胴体角誤差ゲイン，胴体角速度ゲイン，車輪角速度ゲイン，車輪角速度誤差ゲインである

腕のトルクは目標角との差に比例ゲインを掛けたものを目標角速度とする P 制御を行う．

$$\dot{\alpha_d} \quad = \quad k(\alpha_d - \alpha) \tag{9}$$

$\dot{\alpha_d}$ は目標腕角速度，α は腕角，α_d は目標腕角，k は比例ゲインである．

3 評価関数

運動を評価する評価関数を設定する．ここではボールの飛距離が大きいこと以外に，ロボットの安定性や周囲の安全性を評価に含めるため，投球時の移動距離，投球後の胴体角の誤差，移動速度が小さいことが望ましい運動であるとして以下の評価関数を設定した．

$$E = w_1 \cdot l_b^2 - \frac{w_2\int_0^{t_1}(l_r^2)dt}{t_1}$$

$$-\frac{w_3\int_{t_1}^{t_1+t_2}(\theta_{err}^2)dt}{t_2} - \frac{w_4\int_{t_1}^{t_1+t_2}(\dot{\varphi}_{err}^2)dt}{t_2} \tag{10}$$

ここで l_b は飛距離，l_r は投球時のロボットの移動距離，θ_{err} は胴体角の目標値との差，$\dot{\varphi}_{err}$ は車輪角速度の目標値との差，t_1 は投球動作の時間，t_2 は投球動作後に安定性を見る時間，w_1，w_2，w_3，w_4 はそれぞれ飛距離，投球時の移動距離，投球後の胴体角誤差，投球後の車輪角速度誤差に対する重み係数である．今回各モーションに継続時間を学習パラメータに設定しており，一定ではない．このため誤差の積分は時間が増えると大きく，減ると小さくなる．これに対し，飛距離は時間に依存せず，たとえ同程度，もしくはより大きな飛距離を出した場合でも誤差の計算時間によっては評価が逆転してしまう可能性

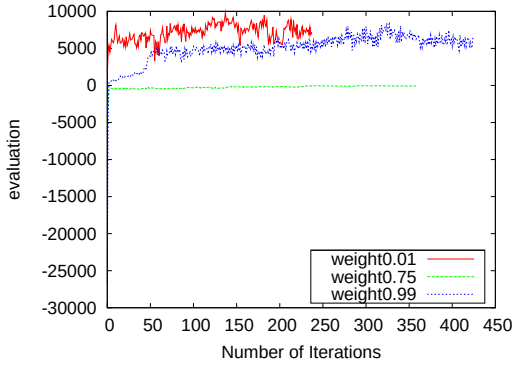
がある．これを回避するため，各誤差の積分をその誤差を計算する時間で割り，誤差の積分の時間平均を評価に利用することにした．

4　実験

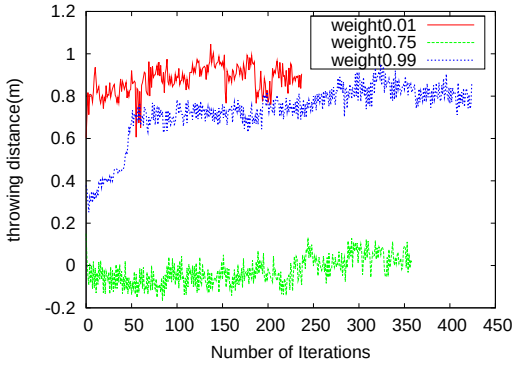
評価の重みによって，学習の進み方が変わることが考えられた．飛距離の重みを 10000 で固定し，1 回目の実験の結果によって他の重みを変更することで 2 回目の学習の進み方がどのように変わるかを確認した．

4.1　移動距離評価重み 1000 での実験

評価の重み w_1，w_2，w_3，w_4 をそれぞれ 10000，1000，100，1 で行ったシミュレーションの結果を示す．各パラメータ初期値から始めた評価，飛距離を **Fig.3** に示す．赤いグラフが軌道重み係数の初期値を 0.01，緑のグラフが 0.75，青いグラフが 0.99 でそれぞれ統一した場合の推移である．



(a) 評価推移



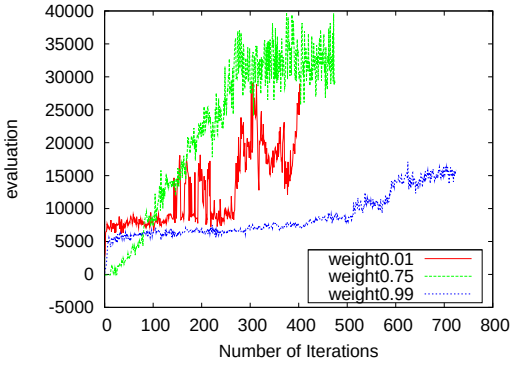
(b) 飛距離推移

Fig.3 移動距離評価重み $w_2 = 1000$ での評価，飛距離の推移

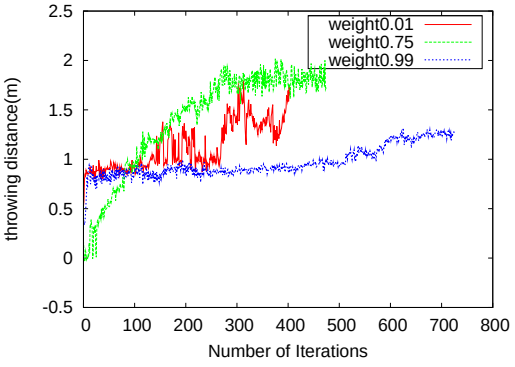
学習の初期に飛距離が正の値になっているにもかかわらず，評価値が大きな負の値になっているのは初期値での姿勢制御が非常に不安定であることを示している．グラフから分かるように数回の学習で評価値は正の値に転じており，初期の姿勢制御の学習には大きな効果があったと言える．しかし，最高でも飛距離 1 m 付近までしか達成できず，飛距離が負の値で始まった場合については 0m 付近から飛距離を伸ばすことができなかったことがグラフから分かる．これは評価の重みから，飛距離の増加による評価値の上昇よりも，安定性の悪化による評価値の減少のほうが大きくなってしまい，より遠くへ投げられるパラメータへ学習が進まなくなったからだと考えられた．

4.2　移動距離評価重み 100 での実験

先の実験結果を踏まえ，評価の要素の内，投球時の動きから評価を出す移動距離の部分が，特に投球姿勢に与える影響が大きいであろうと考え，移動距離の重みを 100 に変更し，投球時の移動距離による負の評価を緩和して再度シミュレーションを行った．つまり w_1，w_2，w_3，w_4 はそれぞれ 10000,100,100,1 である．各パラメータ初期値から始めた評価，飛距離の推移を **Fig.4** に示す．赤いグラフが軌道重み係数の初期値を 0.01，緑のグラフが 0.75，青いグラフが 0.99 でそれぞれ統一した場合の推移である．



(a) 評価推移



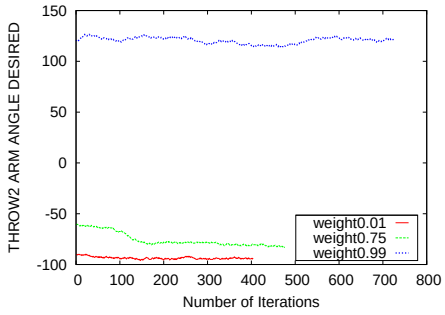
(b) 飛距離推移

Fig.4 移動距離評価重み $w_2 = 100$ での評価，飛距離の推移

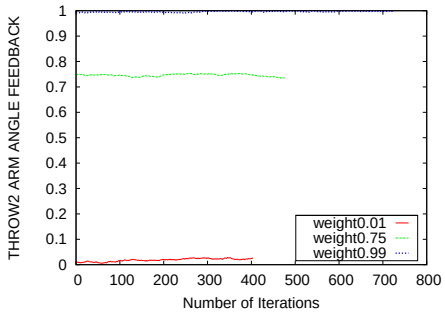
全てのパラメータ初期値の場合で飛距離が 1 m を超え，学習の進み方が移動距離評価重み 100 の場合に比べ促進されたことが分かる．飛距離は最大で 2m を達成している．青いグラフ初期軌道重み係数 0.99 の場合については動きが緩やかで，パラメータの変化による飛距離の増加よりも投球終了後の姿勢制御の評価の減少のほうが大きくなり，学習が大きく進まなかったとが考えられる．逆に赤いグラフ初期軌道重み係数 0.01 の場合については動きが急で不安定なため，パラメータの小さな変化で飛距離が大きく変わったり，転倒したりする場合が多く，学習も不安定になったと考えられる．

投球に最も影響が大きいモーションとして THROW2 モーションの軌道生成に関わるパラメータの学習推移のグラフを **Fig.5**, **Fig.6**, **Fig.7** に示す．赤いグラフが軌道重み係数の初期値を 0.01，緑のグラフが 0.75，青いグラフが 0.99 で学習を進めた場合の推移である．

胴体角や車輪角速度の最終目標値が大きく変化してい

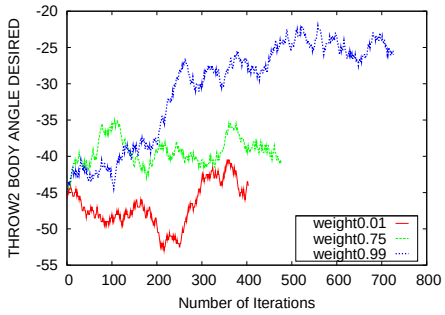


(a) 腕角最終目標値

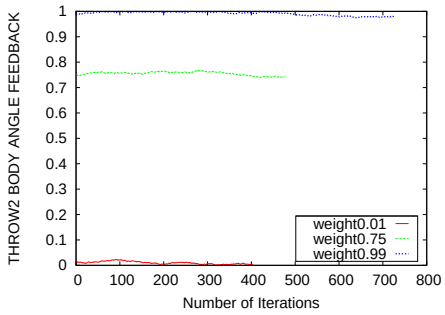


(b) 腕角軌道重み係数

Fig.5 腕角の最終目標値と軌道重みの学習推移



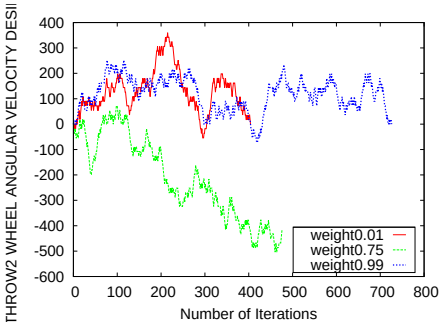
(a) 胴体角最終目標値



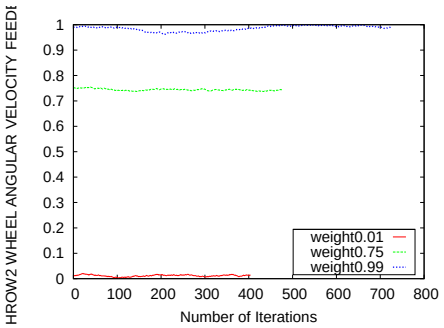
(b) 胴体角軌道重み係数

Fig.6 胴体角の最終目標値と軌道重みの学習推移

るのに対し，腕角の最終目標値や各軌道重み係数はほとんど変化していない．他のモーションにおいても同じ傾向が見られた．このことから，目標軌道の学習は胴体角や車輪角速度の最終目標値を変化させることで進んだことが分かる．



(a) 車輪角速度最終目標値



(b) 車輪角速度軌道重み係数

Fig.7 車輪角速度の最終目標値と軌道重みの学習推移

5 おわりに

倒立二輪型移動ロボットによる投球動作を方策勾配法によって獲得させ，その挙動を動力学シミュレータで調べた．今後の課題として，今回は手動で与えた初期値を自動で探索する，腕を長くするなどより遠くへ投球できるロボットの構造を考察することが考えられる．

参考文献

[1] 妹尾拓, 並木明夫, 石川正俊. 高速スローイング動作におけるエネルギー伝播の解析. 第 7 回計測自動制御学会システムインテグレーション部門講演会 (SI2006)(札幌，2006.12.15)／講演会論文集，pp.736-737, 2006.

[2] 妹尾拓，並木明夫，石川正俊． 波動伝播に基づく高速スローイング動作． ロボティクス・メカトロニクス講演会 2007(ROBOMEC 2007) (秋田, 2007.5.11) ／講演論文集, 1A2-F10, 2007.

[3] 妹尾拓, 並木明夫, 石川正俊. 高速スローイング動作におけるリリース制御の解析. 第 9 回システムインテグレーション部門講演会 (SI2008)(2008 年 12 月 5 日〜7 日・岐阜), 2008.

[4] J.Baxter and P.L.Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Reserch,15:319-350*, 2001.

[5] R.S.Sutton, D.McAllester, S.Singh, and Y.Mansour. Policy gradient methods for reinforcement learning with function approxomation. *Proceedings of the IEEE International Conference on Robotics and Automation(ICRA 2004),pp. 2619-2624*, 2004.

[6] Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. *Proceedings of the IEEE International Conference on Robotics and Automation(ICRA 2004),pp. 2619-2624*, 2004.

[7] 平林友一, 山藤和男. 可変構造型移動ロボットの制御に関する研究 (第 5 報，制御アーム-車輪型モデルによる高速度走行のための慣性力補償). 日本機械学會論文集. C 編 Vol.58, No.552(19920825) pp. 2501-2506, 1992.

[8] 川村伸司. フィードバック制御による倒立ロボットの製作. Interface 特集 はじめての SH-2 基板応用＆開発実践技法 (2006/7) pp. 70-78, 2006.