Actor-Critic 型強化学習を用いたヒューマノイドロボットの動作獲得に関する研究

情報科学科　岩井 優斗　　指導教員：小林 邦和，鈴木 拓央

1　はじめに

ヒューマノイドロボットは身体構造上，人間に似た動作を行うことが可能であり人間社会のなかで人間に代わる役割を担うことが期待されている．またこのためには人間のように多様な動作能力が必要である．時にはシュートモーションといった動作の最適な動作系列を人間が把握できずモデル化ができない動作の獲得も必要である．しかし既存のヒューマノイドロボットの運動制御はモデルベースであり [1]，モデル化できない複雑な動作を獲得することは困難である．そこで動作の獲得に強化学習を用いることで学習主体が試行錯誤によって複雑な動作も学習し人間より優れた動作を獲得する可能性がある [2][3]．

本研究の最終的な目的は人間が明示的なモデルを作成できないような複雑な動作を強化学習により獲得することである．本論文では前段階としてヒューマノイドロボットに単純な動作を強化学習により獲得させることを試みる．

2　Actor-Critic 型強化学習

ヒューマノイドロボットの動作は行動空間が連続であるため，本研究では強化学習の中でも連続な行動空間を扱うことに優れた Actor-Critic 法を用いてモーターにかかるトルクの値を直接的に学習によって獲得する．以下に本論文における Actor-Critic 型強化学習について順を追って説明する．

状態観測

始めに環境より状態 S_t を観測する．

行動

本論文では確率的方策に正規分布を用い平均 μ と分散 σ を更新することで確率的方策を更新する．得られた乱数を直接トルクの値とし，1 ステップ中のモーター角度の変化を行動 a とし時刻 t における状態 S_t から次状態 S_{t+1} へ遷移する．

報酬と強化信号

critic で遷移後の報酬 r_{t+1} を得て，TD 誤差 δ を計算する．TD 誤差は以下のように与える．

$$\delta = r_{t+1} + \gamma V(S_{t+1}) - V(S_t) \tag{1}$$

ここで $V(S_t)$ は状態 S_t の状態価値であり r_t はこのときの報酬である．γ は割引率であり 0 から 1 までの定数である．

確率的方策の更新

TD 誤差に応じて確率的方策を以下のように更新する．

$$\mu_{t+1} = \mu_t + \beta a_t \delta \tag{2}$$

このとき β は学習率であり 0 から 1 までの定数である．ここでは確率的方策の更新として正規分布の平均を更新する．TD 誤差に同ステップの行動を掛けることで TD 誤差が正ならその行動をとる確率を高め，負なら低くする．

状態価値関数の更新

TD 誤差に応じて critic にて TD 誤差が 0 に近づくように状態価値を以下のように更新する．

$$V(S_t) \leftarrow V(S_t) + \alpha\delta \tag{3}$$

以上を毎ステップごとに繰り返し学習を進める．

3　計算機シミュレーション

ここでは 3 次元空間上にヒューマノイドロボット NAO を配置し，直立状態から図 1 のような右手を水平に伸ばしたポジションへの振り上げ動作を学習によって獲得させる．動作のために右肩のピッチにかかるトルクの値を学習によって決定する．この問題設定のためのパラメータを以下のように設定した．μ の初期値を 0.0，σ の初期値を 0.4，α を 0.05，β を 0.01，γ を 0.9 とした．また報酬は式 (4) で示す．

$$r = -\theta + 1.55 \tag{4}$$

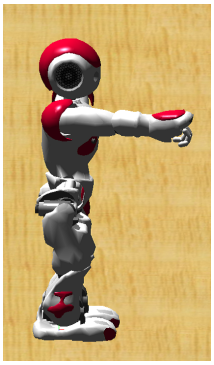

図 1　目標の概観

ここで，θ は目標からの角度をラジアンで表したものであり $(0 \leq \theta \leq 1.55)$ のみ式 (4) で与えその他は報酬 0 とする．

次にシミュレーションの結果として各エピソードの目標への最小到達ステップ数と累積報酬をエピソード数を横軸にとりプロットした．それぞれ図 2，3 に示す．

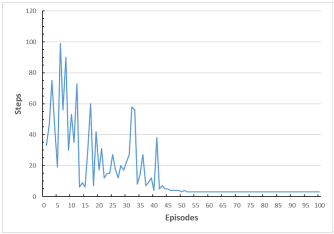

図 2　最小到達ステップ数

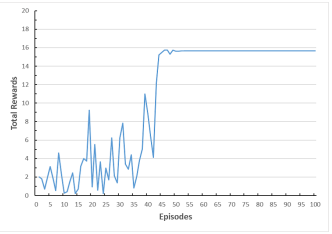

図 3　累積報酬

以上よりエピソード数の増加とともに最小到達ステップ数は小さくなり累積報酬が大きくなりやがて収束していることがわかる．これは学習が終了し，問題設定に対して正しい行動系列を獲得していることがわかる．

4　まとめ

本論文ではヒューマノイドロボットが単純な動作をモデルに頼らず Actor-Critic 型強化学習を用いて獲得することに成功した．このことはヒューマノイドロボットがより複雑でモデル化が困難な動作に対しても Actor-Critic 型強化学習を用いることで獲得する可能性を示している．本研究の今後の展開ではより複雑な動作であるヒューマノイドロボットのシュートモーションを学習させることを考えている．また実機での実装も考え，報酬の遅れに対処すべく適正度の履歴を導入するなど Actor-Critic 型強化学習の改良を行う．

参考文献

[1] 國吉 康夫, 大村 吉幸, 寺田 耕志, 長久保 晶彦: “等身大ヒューマノイドロボットによるダイナミック起き上がり行動の実現”, 日本ロボット学会誌, Vol.23, pp.706–717 (2005)

[2] R. S. Sutton and A. G. Barto (三上 貞芳, 皆川 雅章 共訳): 「強化学習」, 森北出版 (2000)

[3] 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, pp.618–623 (1999)