1R-3

段階的学習による高自由度ロボットの強化学習の効率化

程原　教文†　　中村　真吾††　　橋本　周司†††

早稲田大学先進理工学研究科†　　芝浦工業大学工学部††　　早稲田大学理工学術院†††

1. はじめに

近年，ロボットが自律的に行動を獲得するための方法として，強化学習[1]が注目されている．しかしながら，強化学習にはロボットが取り得る自由度が増大すると，状態空間の大きさが指数関数的に広がり，学習時間が膨大になるという「次元の呪い」と呼ばれる問題がある[2].

この問題を解決するために，我々は段階的学習法を研究している[3]．これは，強化学習を状態空間の探索範囲を限定して行う事前学習と，事前学習で得た学習結果を用いて状態空間全域を学習する発展学習の二段階に分けて行うことで，学習を効率化する手法である．これまでに，本手法を多関節アームロボットのリーチング動作獲得に適用し，シミュレータ上で学習時間を短縮できることを確認している．

本論文では，段階的学習法を改良し，複数のパターンで事前学習を行った結果を合成してから追加学習を行うことで，より複雑なタスクの学習を効率化する手法を提案する．また，提案手法をアームロボットのボール打ち運動の獲得に適用し，提案手法の有効性を検証する

2. 提案手法

2.1. 強化学習

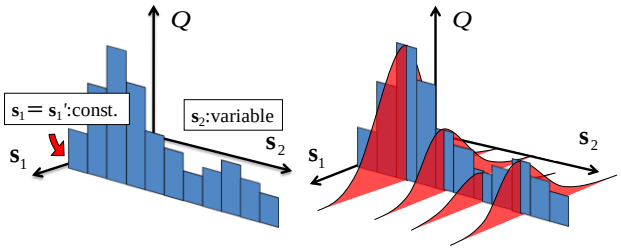
本論文では強化学習の代表的なアルゴリズムである Q-Learning を用いる．これは時刻 t の状態 $\mathbf{s}_t$ で行動 $\mathbf{a}_t$ を取った時，行動価値関数 $Q(\mathbf{s}_t, \mathbf{a}_t)$を式(1)で更新していく手法である．

$$Q(\mathbf{s}_t,\mathbf{a}_t) \leftarrow Q(\mathbf{s}_t,\mathbf{a}_t)+\alpha[r_{t+1}+\gamma \max_a Q(\mathbf{s}_{t+1},\mathbf{a})-Q(\mathbf{s}_t,\mathbf{a}_t)] \quad (1)$$

ここで r_t は時刻 t に得られる報酬，α は学習率，γは割引率である．

2.2. 段階的学習法

段階的学習法は，事前学習と発展学習の二段階で構成される．まず，事前学習では，状態空間の次元を制限し，探索範囲を限定して強化学習を行う．つまり，状態空間の次元数 N のうち n 個の値を固定して学習を行う（図 1(a)）．つぎに，事前学習で得た結果の Q 値を式(2)で状態空間全域に拡張する（図 1(b)）．



(a)事前学習による Q 値　(b)Q 値の拡張

図 1　段階的学習の概要図.

$$Q(\mathbf{s}=(\mathbf{s}_1,\mathbf{s}_2))=C\cdot Q(\mathbf{s}_1',\mathbf{s}_2)\exp\left[-\frac{(\mathbf{s}_1-\mathbf{s}_1')^2}{\sigma^2}\right] \quad (2)$$

ここで $\mathbf{s}_1$，$\mathbf{s}_2$は固定した，又は固定しなかった状態変数ベクトル，σ はガウス関数の標準偏差，C は補正定数である．発展学習では，この拡張された Q 値を初期値として，状態空間全域を使った学習を行うことで，最適な Q 値が獲得される．

2.3. 複数の事前学習の合成

ここでは複数の次元制限パターンで行った事前学習を図2のように合成することで，段階的学習法を拡張する手法を提案する．まず，事前学習を状態空間の次元の制限パターンを変えて行い，複数の異なるパターンのQ値を得る．次に式(3)のように，各パターンのQ値を状態空間全域に拡張した後，合成する．

$$Q(\mathbf{s}=(\mathbf{s}_1,\mathbf{s}_2,...))=\sum_i C\cdot Q(\mathbf{s}_i,\mathbf{s}_{j\neq i})\exp\left[-\frac{(\mathbf{s}_i-\mathbf{s}_i')^2}{\sigma^2}\right] \quad (3)$$

ここで$\mathbf{s}_i$は事前学習にて固定する状態ベクトルである．このQ値を用いて追加学習を行うことで，より高い自由度を持つロボットであっても，学習にかかる時間を短縮できることが期待できる．

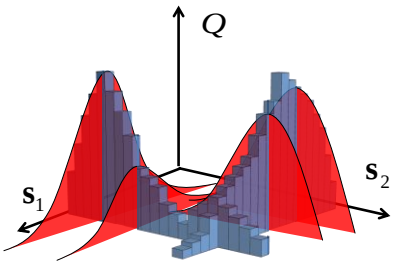


図 2 事前学習の合成.

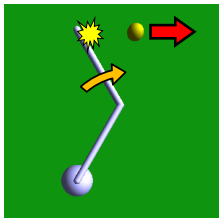


図 3 アームロボット.

"Efficient Reinforcement Learning by Phased Approach for High DOF Robots" Norifumi Hodohara†, Shingo Nakamura††, Shuji Hashimoto†††
Graduated Schools of Advanced Science and Engineering, Waseda University†
College of Engineering, Shibaura Institute of Technology††
Faculty of science and engineering, Waseda University†††

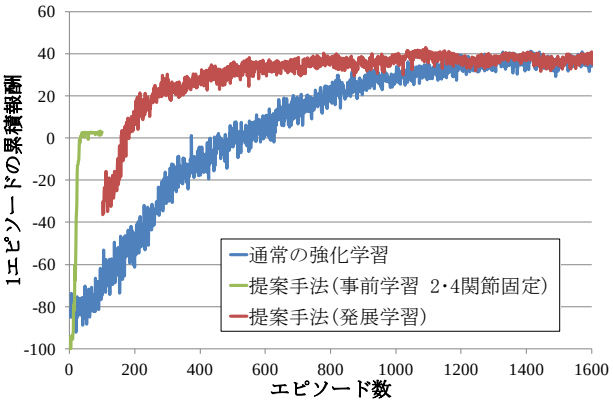
3. 評価実験

提案手法の有効性を検証するために，多関節アームロボットに提案手法を適用して実験を行った．シミュレータ上に 4 関節及び 5 関節アームロボットを作製し，4 関節アームロボットには段階的学習法を，5 関節アームロボットには事前学習の合成手法を適用し，エピソードごとの累積報酬の推移を記録する．学習させるタスクは，図 3 のように静止しているボールを特定の方向へ出来るだけ強く打ち飛ばすこととした．

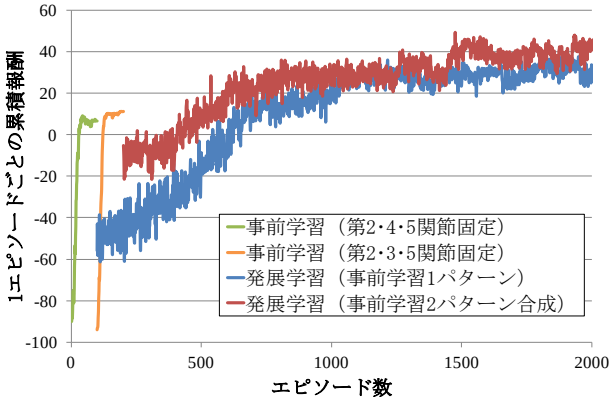
アームロボットの各関節(可動域は–90°~90°)を 30°刻みで分割した角度を状態 **s**，各関節を時計回り又は反時計回りに 30°回転させる動作を行動 **a** として Q-Learning を適用する．タスクを達成できた場合は報酬 r として打ち飛ばしたボールの速さの 2 乗値を与え，間違った方向へ打ち飛ばした場合は–100 の罰則を与える．行動方策の選択則としては SoftMax 法を用いた．学習率と割引率は，4 関節アームではα=0.6，γ=0.8 を，5 関節アームでは α=0.6，γ=0.9 を用いた．

この場合，状態 **s** の次元数はアームの関節数と等しい．事前学習では，アームロボットのいくつかの関節を固定して状態空間の次元を制限し，発展学習では全関節を動作させることで段階的学習法を行う．4 関節アームロボットでは第 2･4 関節を事前学習で固定した．5 関節アームロボットにおいては，第 2･4･5 関節及び第 2･3･5 関節固定の 2 パターンの事前学習を用意した．事前学習で得た Q 値を拡張するガウス関数の標準偏差σと補正定数 C は，4 関節アームではσ=0.3，C=2.0，5 関節アームではσ=0.2，C=1.0 とした．

図 4(a)に 4 関節アームロボットに段階的学習法，および通常の Q-Learning を適用した結果を，図 4(b)には 5 関節アームロボットに事前学習 1 パターンのみの Q 値，又は 2 パターンを合成した Q 値を追加学習に利用した結果を示す．それぞれ累積報酬を比較した．なお実験結果は 100 回の試行の平均である．図 4(a)では，提案手法を適用した場合は通常の強化学習よりも累積報酬の上昇が早まり，学習の収束が早まっていることが確認できる．発展学習の前に必要な事前学習も短時間で収束している．図 4(b)では，事前学習 2 パターンを合成した場合の発展学習は，事前学習 1 パターンのみしか用いなかった場合よりも学習初期での獲得報酬が高くなっているが，学習の収束速度に関してはあまり有意な差は確認できなかった．これは，2 つの事前学習パターンの状態空間に存在する位置がそこまで大きく異なっていないために，提案手法の効果があまり現れなかったためと考えられる．



(a) 4 関節アームロボット



(b) 5 関節アームロボット

図 4 エピソードごとの累積報酬の推移.

4. まとめと今後の課題

本論文では，「次元の呪い」問題を解決するために段階的学習法を検討し，複数パターンの異なる事前学習を合成することで複雑なタスクの学習を効率化する手法を提案した．提案手法を多関節アームロボットのボール打ち飛ばし動作獲得に適用し，提案手法の有効性を確認した．今後はアルゴリズムの改良に取り組むと共に，提案手法をアームロボットの動作獲得以外のタスクに適用し，提案手法の応用範囲を広げたい．

謝辞

本研究の一部は，早稲田大学ヒューマノイド研究所，グローバルCOE プログラム「グローバルロボットアカデミア」の研究助成を受けて行われた．

参考文献

[1] R. S. Sutton and A. G. Barto "Reinforcement Learning: An Introduction." Cambridge, MA: MIT Press, (1988).

[2]伊藤一之, 松野文俊,, 五福明夫, "強化学習による冗長ロボットの自律制御に関する研究 -身体像を考慮した強化学習-" 日本ロボット学会誌, Vol. 22, No. 5, pp.130-147, 2004.

[3] Norifumi Hodohara, Yuichi Murakami, Shingo Nakamura, Shuji Hashimoto "Reinforcement learning with phased approach for fast learning", Proceedings of the International Symposium on Artificial Life and Robotics (AROB 17th '12), pp.930-933, (2012).