2N5-OS-03b-1

LS-Q 学習による探索と停滞ループの回避

Exploration and Stagnant Loop Avoidance by LS-Q Learning

浦上 大輔[*1] Daisuke Uragami
高橋 達二[*2] Tatsuji Takahashi
高橋 優太[*2] Yuta Takahashi
アルアル ワン アリー[*3] Ali Alalwan
松尾 芳樹[*1] Yoshiki Matsuo

[*1] 東京工科大学コンピュータサイエンス学部 School of Computer Science, Tokyo University of Technology
[*2] 東京電機大学理工学部 School of Science and Technology, Tokyo Denki University
[*3] 東京工科大学大学院 バイオ・情報メディア研究科 Tokyo University of Technology Graduate School

In a previous study, we proposed a novel reinforcement learning architecture LS-Q that applies human cognitive biases to action selection in Q-learning. It has become clear that LS-Q learning adaptively searches in the environment with large uncertainty. In this study, we analyze the aspect in which LS-Q learning adeptly escapes from local optima, avoids stagnant loops through states with little rewards, and return to efficient motion learning.

1. はじめに

環境との相互作用に基づく学習アーキテクチャーとして強化学習が注目されている. 強化学習では, 未知の環境において, 試行錯誤的に探索しつつより多くの報酬を得ることを目的の1つとしている. 有限回の試行錯誤において, 探索によって環境の知識を蓄積することを優先するか, 既に得られている知識に従って報酬を獲得することを優先するかの判断は難しい. くわえて, 報酬の遅延が探索と知識利用の選択をより困難にする. このような課題をふまえて, 我々は, 人間の推論傾向(論理階層を混同する傾向)を模倣して Q 学習に応用する強化学習アルゴリズム, LS-Q を提案している[Uragami 2014]. LS-Q 学習は, 不確実性の大きい環境において適応的に探索を行うということが明らかになりつつある.

本研究では, 大車輪ロボットの運動獲得を例として, LS-Q 学習が報酬の少ない状態でのループ(停滞ループ)を巧みに回避する様相を解析し, 普遍的な探索理論における局所性や論理階層の混同[高橋 2013] あるいは内部観測[松野 2000] の意義と効用を考察する.

2. LS-Q 学習による大車輪ロボットの実現

人間の認知バイアス(推論や意思決定における偏り)の1つとして, 「p ならば q」より「q ならば p」を推論する傾向性, 対称性バイアスがある. このような推論は論理的には必ずしも正しくないが, 経験的にはしばしば有用であり, 人間の知能の柔軟さに関係していると考えられる[服部 2008]. LS モデルは, 人間の対称性バイアスを定量的に再現する一方で, 1 状態の強化学習課題である n 本腕バンディッド問題において優れたパフォーマンスを示すことが知られている[篠原 2007].

我々は, 状態数が N 個であるより一般的な強化学習課題に適応可能な強化学習手法である Q 学習の行動選択に, LS モデルを応用する手法, LS-Q 学習を提案している. Q 学習は, 実装が容易で多くの学習対象に適用可能であるが, 学習に要する試行回数が膨大であることや環境の不確実性の影響が大きいなどの課題がある. LS-Q 学習は, 学習速度を改善し且つ環境の不確実性の影響を小さくすることにより, 大車輪ロボットの制御(シミュレーション)に成功している(図 1).

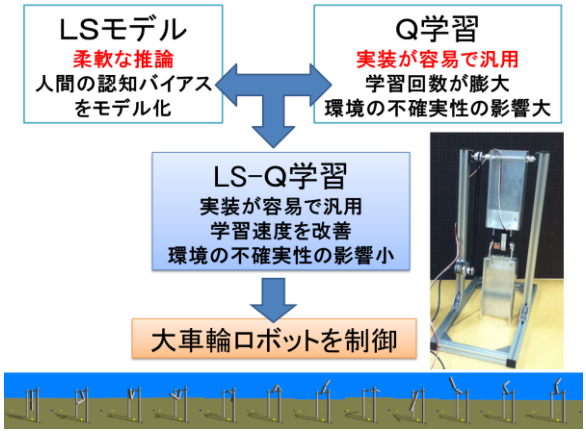

図 1 : LS-Q 学習による大車輪ロボットの制御

2.1 LS-Q 学習

LS-Q 学習の行動選択と学習のアルゴリズムを簡単に述べる(図 2). 詳細は文献[Uragami 2014]による. LS-Q 学習では, C-table に基づいて行動選択を行う. C-table とは, greedy な行動(Q 値が最大の行動)を選択した回数とそれ以外の行動を選択した回数を, 状態ごとに記録したものである. C-table を次式のように LS モデルで評価する.

$$\mathrm{LS}(\text{greedy}|\text{行動 A}) = \frac{a + bd/(b+d)}{a + bd/(b+d) + b + ac/(a+c)} \quad (1)$$

連絡先:浦上大輔, 東京工科大学コンピュータサイエンス学部, 東京都八王子市片倉町 1404-2, uragami@cs.teu.ac.jp

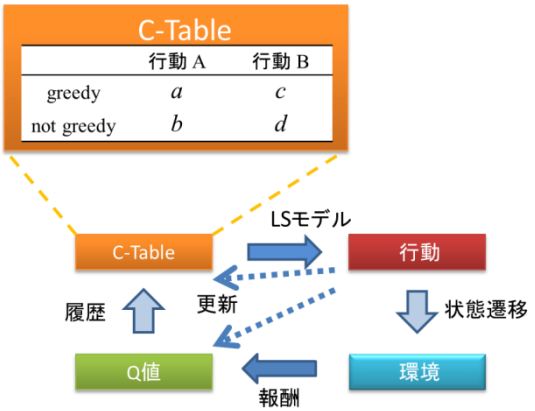

図 2：LS-Q 学習の行動選択と学習のアルゴリズム

LS(greedy|行動 A) と LS(greedy|行動 B)を計算し，値の大きい方の行動を LS-greedy な行動とする．学習の過程では，ε-greedy 行動選択によって，LS-greedy な行動とそれ以外の行動を定められた確率で選択する．Q 値の更新は通常の Q 学習と同様の手続きで行う．

2.2 大車輪ロボットへの適用

大車輪ロボットとは図 1 中の写真のように体操の鉄棒競技を模したロボットである．鉄棒とロボットの接続部分はフリージョイントになっており，腰部の関節のみが能動的に稼働することによって振り上げ運動を行う．鉄棒とロボットの接続部分を第1ジョイント，腰部の関節を第2ジョイントと呼ぶことにする．第1ジョイントの角度と角速度，第2ジョイントの角度をそれぞれ分割して離散化することにより，LS-Q 学習／Q 学習における状態を定義した．行動は，第2ジョイントを曲げる A0，伸ばす A1，停止 A2 の3通りである．報酬は，鉄棒に対するつま先（第2ジョイントの先端）の振り上げ角度に比例するように定義し，つま先が鉄棒の真上に位置する場合に 1（最大値），つま先が鉄棒の真下に位置する場合に 0（最小値）となるように設定した．本研究では，ODE (Open Dynamics Engine) を用いたシミュレータを構築して実験を行った．

3. 停滞ループの回避

これまでの研究成果として，状態分割が粗く不確定性が大きい環境において，LS-Q 学習は Q 学習より良いパフォーマンスを示すことが明らかになっている[Uragami 2014]．これをふまえて本研究では，状態分割を極端に粗くし，第1ジョイントの角度を 6 状態（P0, P1, P2, P3, P4, P5），第1ジョイントの角速度を 3 状態（W0, W1, W2），第2ジョイントの角度を 3 状態（R0, R1, R2），合計 6×3×3＝54 状態とした．状態数が少ないと不確定性が大きくなるが，学習時間を短縮できるという実用上の利点にくわえて，状態遷移が可視化し易くなるという解析上の利点がある．

図 3 は，シミュレーション結果（100 試行の平均）である．0.2 秒を 1 ステップとして行動選択と Q 値および C-table の更新を行う．横軸は学習時間で，1000 ステップを 1 セットとしてロボットの状態を初期状態に戻している．縦軸は，1 セット毎の獲得報酬の合計である．学習の初期では完全にランダムに行動を選択し，1 セット毎に LS-greedy／greedy な行動を選択する確率を 0.02 増やした．最後の 10 ステップは完全に LS-greedy／greedy な行動を選択する．LS-Q 学習と Q 学習を比較すると，学習の開始から徐々に LS-Q 学習の獲得報酬が Q 学習より上回っていることがわかる．また，Q 学習では学習の終盤，ランダムな行動選択の割合が小さくなると，獲得報酬が減少している．一方，LS-Q 学習では，獲得報酬が一旦は減少しているが，その後回復して学習の終了時には高い獲得報酬を得ている．

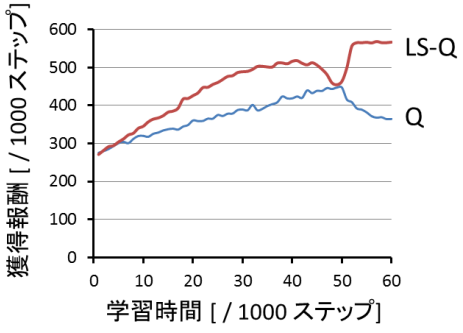

図 3：学習曲線

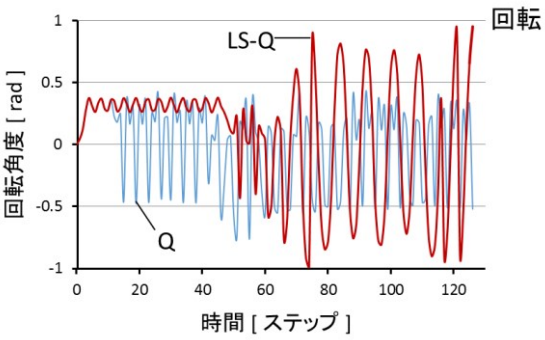

図 4：回転角の変化

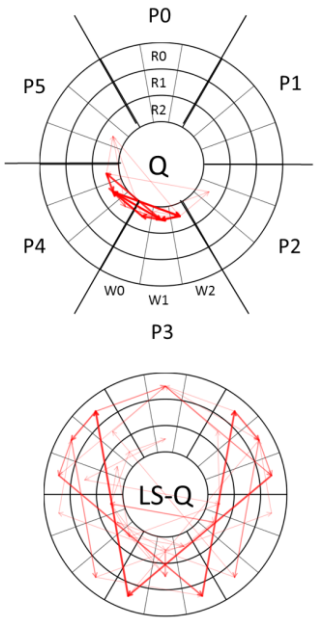

図 5：状態遷移図

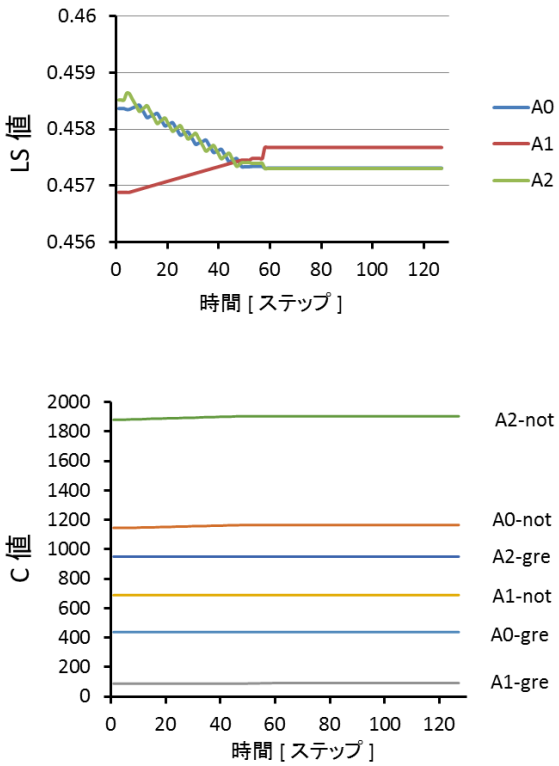

図 6：LS-Q における LS 値と C 値の変化

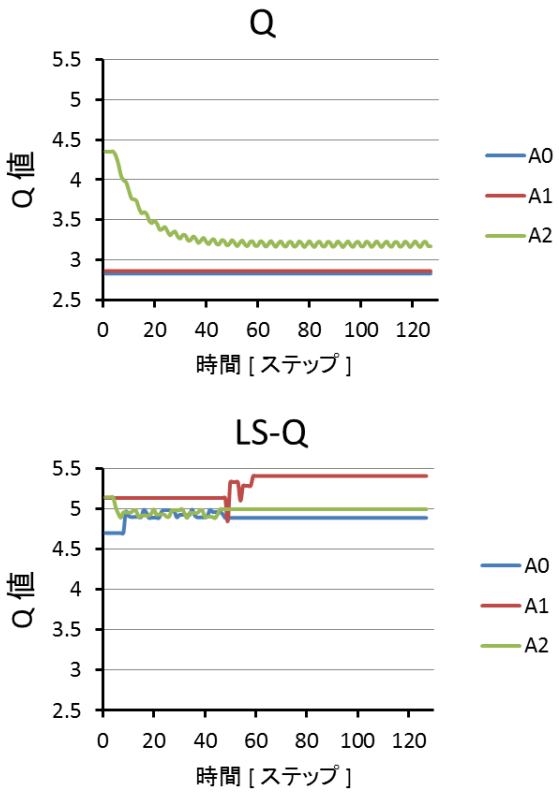

図 7：Q と LS-Q における Q 値の変化

学習の終盤で獲得報酬が減少する現象は，報酬の少ない状態でのループ（停滞ループ[坂井 2010]）が原因である．図 4 は，ランダムな行動選択の割合が 0 になる 51 セット目の最初の 127 ステップについて，第1ジョイントの回転角の変化をプロットしたものである．Q 学習では，回転角が 0.5 ラジアンから−0.5 ラジアンの間で振動していることがわかる．一方，LS-Q 学習では，停滞状態（5〜45 ステップ付近）から過渡状態（45〜70 ステップ付近）を経て回転に至っている．

図 5 は，52 ステップから 127 ステップまでの状態遷移図である．同心円を 54 分割し，それぞれの区画に（P0, P1, P2, P3, P4, P5）×（W0, W1, W2）×（R0, R1, R2）で指定される1つの状態が割り当てられている．直線は2つの状態間の遷移を表し，遷移回数が多いほど線を太くしている．Q 学習（上図）の場合，同心円の中心側（R2）の P3 と P4 間に遷移が集中している．これは第2ジョイントを曲げた姿勢で，第1ジョイントの回転角が小さい状態に留まっていること（停滞ループ）を意味している．より多くの報酬を得るためには，一見して不適切あるいは遠回りに見えても，一旦は第2ジョイントを伸ばして報酬が小さい状態を経由する必要がある．つまり，いわゆる報酬の遅延が問題となる．Q 学習は，原理的には（環境がマルコフ性を満たすなどの条件の下で）報酬の遅延を換算してより多くの報酬を得るように行動選択を最適化する．しかし，本研究の学習環境は状態分割が粗く環境の不確実性が大きいため，最適な行動を得ることができていない．一方，LS-Q 学習(図 5 下)では，第2ジョイントを伸ばした姿勢（R0）を経由して，第1ジョイントの回転角が大きい状態（P0, P1, P5）に至っている．つまり，LS-Q 学習は停滞ループを回避している．

LS-Q 学習はどのようなメカニズムで停滞ループを回避しているのだろうか？図 6 は，停滞ループ中のある状態（P3, W1, R2）の LS 値（上図）および C 値（下図），すなわち C-table 上の値，の変化である．この状態は第2ジョイントを曲げてつま先を振り上げた状態（R2）であるため，行動 A1（伸ばす）を選択してつま先を振り下げる必要がある．しかし，上図の初期のステップでは A0 または A2 の LS 値が最大となっている．このため，ロボットはこの状態に留まる．しかし，A0 と A2 の LS 値は降下する一方で，選択されていないにも関わらず，相対評価[高橋 2013]を通じて A1 の LS 値が上昇し，A1 の LS 値が A0 と A2 の LS 値を逆転する．その結果，A1 が選択されて停滞状態から脱出する．ここで注目すべきことは，C 値（下図）はほとんど変化していないことである．図中の A1-gre や A2-not はそれぞれ，「A1 を選択してその行動の Q 値が最大であった回数」や「A2 を選択してその行動の Q 値が最大でなかった回数」である．図中の 0 ステップの段階で既に，1000 ステップ×50 セットの学習時間を経ている．そのため，0〜40 ステップあたりまでは連続して A0 および A2 が選択され A0-not および A2-not が増加しているが，比率として僅かである．また，その間 A1 は選択されていないため，A1-gre および A1-not は変化していない．にもかかわらず，LS 値（上図）は急激に変化している．これは正に LS モデル（式(1)）の効果であり，人間の推論の柔軟さの効用であると考えられる[高橋 2013]．

図7は，通常の Q 学習（上図）と LS-Q 学習（下図）の Q 値の変化である．通常の Q 学習では A2 の Q 値が最大となっており，A2 が選択され続けることにより A2 の Q 値は減少しているが，他と逆転するに至っていない．一方，LS-Q 学習では，前半は複数の行動が入れ替わり最大値となっているが，後半は A1 が最大値となり，Q 値においても適切に学習されていることがわかる．通常の Q 学習と LS-Q 学習のどちらも，この図の Q 値に至るまでに十分な回数の Q 値の更新を行っているが，適切な Q

値に収束していない. その原因は環境の非マルコフ性にあると考えられるが詳細な解析は今後の課題である.

4. おわりに

本研究では人の認知特性を応用した強化学習アルゴリズムが, 探索効率を高め且つ停滞ループを回避する様相を解析した. 停滞ループの要因は環境の不確実性, 非マルコフ性にあると考えられる. 非マルコフ的な環境において学習可能な強化学習アルゴリズムはいくつか提案されている. その基本は潜在するダイナミクスを再構成することにあり, そのためには時系列を十分に記憶し, 且つその時系列を見渡す別のサブシステムが必要である. 工学的にはそのような方法も1つの選択肢であろう. しかし, 人間の認知特性に触発されたアーキテクチャーの開発あるいは人間の知能の理解において問われるべきは, このような記憶の起源であり, サブシステム(=脳の中の小人)の不在あるいは創発である. 太田は,「1 ステップの計算の中に内包される過去の部分的な(真の意味での並列な)再構築」を人工知能あるいは脳科学における最も重要な問題として掲げている[太田 2014]. 本研究の文脈において太田の主張をパラフレーズすると,「非マルコフ的な環境におけるダイナミクス(=1 ステップの計算の中に内包される過去)のサブシステムなしの(=真の意味での並列な)部分的な再構成」となるであろう. 本研究の成果によると, LS モデルすなわち人間の対称性バイアスは, 上記の再構成あるいは“再現前化” [Deleuze 1968]に関係していると考えられる. このような視点から再現前化=表象あるいは記号を再考することは, 記号創発ロボティクス[谷口 2000]と内部観測を繋ぐ/切断する補助線となるであろう.

謝辞

本研究の一部は平成 25 年度東北大学電気通信研究所 共同プロジェクト研究 H25/A12「不定な環境における適応能の階層横断的解明と工学的応用」, 東京電機大学総合研究所研究 Q13K-03, Q11K-02, 日本学術振興会科学研究費補助金 25730150 による.

参考文献

[Uragami 2014] Uragami, D., Takahashi, T., Matsuo, Y.: Cognitively inspired reinforcement learning architecture and its application to giant-swing motion control, *BioSystems*, 116, 1-9, 2014.

[高橋 2013] 高橋達二: 論理的階層の圧縮としての内部観測によるトレードオフの乗り越え, 2013 年度人工知能学会全国大会(第 27 回)予稿集, 1L3-OS-24a-2, 2013.

[松野 2000] 松野孝一郎: 内部観測とは何か, 青土社, 2000.

[服部 2008] 服部雅史: 推論と判断の等確率性仮説:思考の対称性とその適応的意味, 認知科学, 15, 3, 408–427, 2008.

[篠原 2007] 篠原修二, 田口亮, 桂田浩一, 新田恒雄: 因果性に基づく信念形成モデルと N 本腕バンディット問題へ応用, 人工知能学会論文誌 22 巻 1 号 G, pp.58-68, 2007.

[坂井 2010] 坂井直樹, 川辺直人, 原正之, 豊田希, 薮田哲郎: 強化学習を用いたスポーツロボットの大車輪運動の獲得とその行動形態の考察, 計測自動制御学会論文集 Vol.46, No.3, pp.178-187, 2010.

[太田 2014] 太田宏之: マルコフ性を前提としない情報処理に向けて, 第 25 回計測自動制御学会 SI 部門共創システム部会研究会・第 8 回内部観測研究会(共同開催), 2014.3.8 口頭講演(予定).

[Deleuze 1968] Deleuze, G., Différence et repetition, P.U.F., 1968 (『差異と反復』財津理訳, 河出書房新社, 1992).

[谷口 2010] 谷口忠大: コミュニケーションするロボットは創れるか―記号創発システムへの構成論的アプローチ, NTT 出版, 2010.