2Q-09

非対称学習率による報酬確率分布の弁別性向上

Improved discrimination of reward probability distribution by asymmetric learning rate

佐鳥 玖仁朗 † Kuniaki Satori　吉田 豊 † Yutaka Yoshida　神谷 匠 † Takumi Kamiya　太田 宏之 ‡ Hiroyuki Ohta　高橋 達二 † Tatsuji Takahashi

東京電機大学 † Tokyo Denki University　防衛医科大学校 ‡ National Defense Medical College

1. 序論

動物が生息する環境においては一般に，食物は必ずしも同じ場所から毎回得られるわけではない．そのような過酷な環境における探索行動の際，良い経験と悪い経験について非対称な学習を行う可能性が示唆されている．

探索行動を説明するモデルとしては，これまで Q 学習モデルをベースとした様々な強化学習モデルが提案されている．その中の一つに，非対称な学習方法を組み込んだモデルとして，Dual learning rate Q (DLR-Q) がある [1, 2, 3]．一般的に Q 学習では報酬予測誤差に対して単一の学習率を用いて学習を行う [4] が，DLR-Q は報酬予測誤差の正負によって異なる学習率を用いる．我々は，5 本腕バンディット問題を実行するマウスの行動系列を DLR-Q モデルにフィッティングさせて解析したところ，マウスは環境の報酬に応じて非対称な学習を行っているとの推定を得ることができた．

先行研究によって，DLR-Q には選択肢間の弁別性を上げ，獲得報酬を増加させるという性質があることが確かめられており，マウスの非対称な学習には合理性があるものと考えられる．この DLR-Q の性質の詳細を調べることで，報酬分布に応じた最適な学習戦略を導出でき，性能向上につながる可能性がある．しかし，先行研究はベルヌーイ型のバンディット問題を対象とした解析であり，報酬確率分布の平均と分散に差がある場合に関する一般的な解析が行われていない．そのため本研究では，報酬確率が正規分布となる環境下で DLR-Q の性質を分析した．

2. 非対称な強化学習

動物やヒトは，良い経験と悪い経験の両方から学ぶ．成功した行動は強化され，失敗した行動は抑制されることでより適切な判断ができるようになる．そのような学習は Q 学習モデルで説明がされ，成功と失敗に対して同じ学習率を用いて学習を行っている．しかし，必ずしも良い経験と悪い経験の両方から等しく学習するわけではないと考えられる．ヒトを対象とした研究では，ドーパミン補充薬の投与やドーパミンシグナル伝達経路関連遺伝子の多型によって，パーキンソン病患者における正と負の学習率が変化することが報告されている [1, 5, 6]．また，我々のマウス行動課題実験の結果から，マウスは報酬確率が低い場合，負の学習率に比べて正の学習率が高くなる傾向が示された．

本研究において DLR-Q は以下のように定義される．ある行動 a_t が選択され，報酬 r_t が与えられた各試行 t ごとに，行動価値 Q_t を以下の式 (1) ように更新される．

$$\begin{aligned} Q_{t+1}(a_t) &= Q_t(a_t) + \begin{cases} \alpha^+ \Delta Q_t & \text{if } \Delta Q_t \geq 0 \\ \alpha^- \Delta Q_t & \text{if } \Delta Q_t < 0 \end{cases} \quad (1) \\ \Delta Q_t &= r_t - Q_t(a_t) \end{aligned}$$

連絡先: 高橋達二, 東京電機大学理工学部, 埼玉県比企郡鳩山町石坂, Tel: 049-296-5416, tatsujit@mail.dendai.ac.jp

α^+ と α^- は，それぞれ選択された行動に対する正と負の学習率を表し，報酬予測誤差が正である際には α^+ で更新され，負である際には α^- で更新される．

DLR-Q が分析されている先行研究では，ベルヌーイバンディット問題を用いて DLR-Q の非対称な学習率に関しての理論的解析を行っている．解析によると，DLR-Q は正と負の学習率の比率 $x(=\alpha^+/\alpha^-)$ に応じて価値の過大評価・過小評価を行う．環境の報酬分布に対して適切な比率 x は，価値間の差を広げることで選択肢の弁別性を上げ，獲得報酬の増加につながる．

しかし，先行研究の分析では報酬分布として 0, 1 の 2 値しか存在しないベルヌーイ分布のみを対象としていた．ベルヌーイ分布は報酬確率となる平均によって分散も決定されるため，報酬の平均と分散を分離した分析は行われていない．そこで本研究では，より一般的な連続報酬量である正規分布を報酬に用いることで平均と分散を独立に変更して分析を行う．

3. 実験

報酬に正規分布を用いたバンディット問題により平均と分散に関しての DLR-Q の性質を分析する．実験は，選択肢の報酬確率分布の平均が異なり分散が等しい場合，平均が等しく分散が異なる場合の二つのバンディット問題を想定し行う．分析方法として，正負の学習率比 x について 1 を中心とした複数の x で実験を行うことに よりその性質を比較した．この $x = 1$ である場合は通常の Q 学習と一致する．方策は softmax を用いて，温度パラメータ $\beta = 1.0$ とした．

3.1 選択肢の平均が異なる実験

平均が 1.0，2.0 の二つの正規分布を用いてバンディット問題を扱う．二つの選択肢のそれぞれを a_1, a_2 とし，この設定では平均=2.0 である a_2 を選択することが最適である．二つの選択肢の標準偏差 σ を同時に変化させながら，各 σ による正答率と選択肢間の価値の差分の比較，Q 値の分布の比較を行う．正答率と選択肢間の価値の差分の比較では，3,000 試行を 1 シミュレーションとし，1,000 シミュレーションの平均を結果とした．Q 値の分布の比較では，各選択肢に対して 100,000 試行の更新を行い Q 値の頻度をヒストグラムとして出力した．結果を図 1，2 に示す．

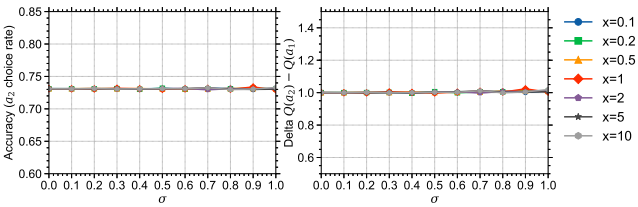

図 1: 二つの選択肢で共有した σ による各 x の正答率と選択肢の価値の差分

縦軸を選択肢間における価値の差分としたグラフ（図 1 右）は，各 σ の設定時での最終的な選択肢間の Q 値の差分を示し

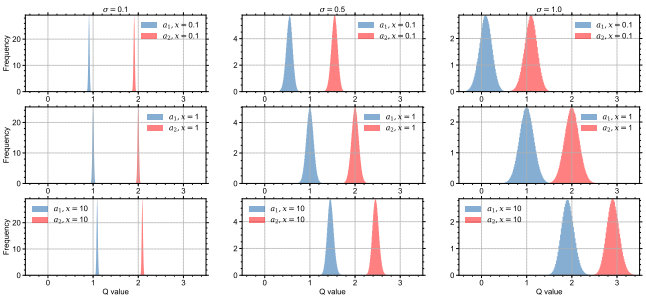

図 2: 二つの選択肢で共有した σ による各 x の Q 値の分布

ている．このグラフから，σ の大きさに限らず選択肢間の価値の差分の大きさは一定であることがわかる．同様に左の正答率も変化していない．また，図 2 は各 σ と各 x における Q 値の分布を示しており，ここから Q 値の分布は σ が大きいほど x による過大評価・過小評価の影響が強く，二つの選択肢の σ が等しいため分布が同時に平行移動していることがわかる．

3.2 選択肢の分散が異なる実験

平均が同じ 1.0 の二つの正規分布を用いてバンディット問題を扱う．a_1 を $\sigma = 0.5$ で固定し，a_2 の σ を変化させながら，各 $\sigma(a_2)$ による結果の比較を行う．結果を図 3，4 に示す．

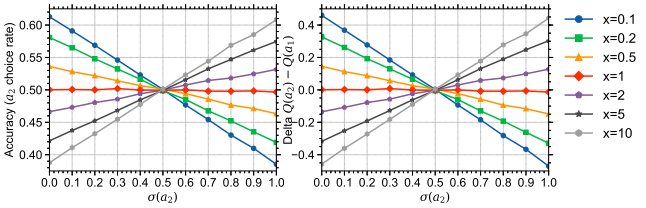

図 3: 二つの選択肢で異なった σ による各 x の a_2 の選択率と選択肢の価値の差分

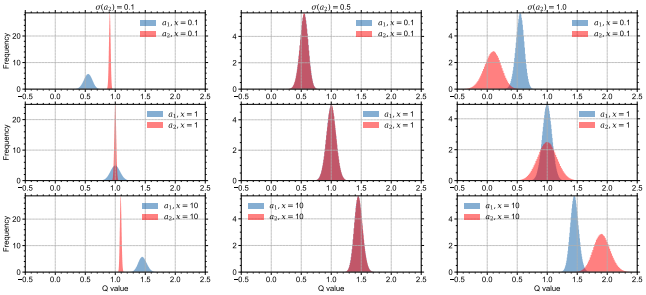

図 4: 二つの選択肢で異なった σ による各 x の Q 値の分布

縦軸を選択肢間における価値の差分としたグラフ（図 3 右）を確認すると，$x = 1$ であるとき σ が異なっても二つの選択肢の価値は均等であることがわかる．対して，$x > 1$ であるとき報酬分布の分散の高い選択肢の価値が高く，$x < 1$ であるとき分散の低い選択肢の価値が高くなることがわかる．縦軸を a_2 の選択率としたグラフ（図 3 左）を確認すると，価値に対応して選択確率が増加・減少している．図 4 の Q 値の分布の結果を確認すると，二つの選択肢の σ の差により Q 値の分布に差が生じ，価値の大小関係ができている．

4. 考察

DLR-Q は実験 3.1 から，x によって Q 値が過大評価・過小評価されており，その傾向は分散が大きいほど強くなっている．そのため，分散が同じ選択肢では同じだけの過大評価・過小評価がなされ，価値の差分は変化せず弁別性の向上にはつながらなかった．また，実験 3.2 から二つの選択肢の平均が等しくとも分散が異なると，x により選択肢間に価値の差が生じ選択傾向が変化した．$x > 1$ であるとき分散の高い選択肢の価値が上昇し分散の高い選択肢を選ぶリスク選好傾向．$x < 1$ であるとき分散の高い選択肢の価値が減少し分散の低い選択肢を選ぶリスク回避傾向がみられた．

先行研究のベルヌーイバンディット問題において適切な x の値が選択肢の弁別性の向上に貢献した理由は，ベルヌーイ分布では平均の大きさによって分散が決定されるため，選択肢で異なった分散によって x の過大評価・過小評価の強さの差が発生し，適切な x を設定することで選択肢間の差分を広げ弁別性の向上につながっていたためと考えられる．

動物の意思決定と比較すると，報酬分布の分散を考慮してリスク態度を変更する性質は先行研究に示されているハチドリの結果に類似している [7]．ハチドリは食物が少ない環境では報酬分布の分散が大きい方の選択を行うというリスク選好性，食物が豊富な環境では分散が小さい方の選択を行うというリスク回避性を示している．DLR-Q の結果と比較すると，食物が少ない環境では $x > 1$ として分散の大きい選択を行い，食物が豊富な環境では $x < 1$ として分散が小さい選択を行うというのに対応している．また，我々のマウスの 5 本腕バンディット問題での解析結果でも，食物が少ない環境で $x > 1$，食物が豊富な環境で x が 1 に近づく傾向が確認されている．このことから，各選択肢の報酬確率分布の分散が異なる場合，非対称な学習により価値の差を拡大もしくは縮小させることでリスク選好性もしくはリスク回避性を示す場合があることが示唆される．

5. 結論

本研究では，非対称な学習を持つ DLR-Q の効用を明らかにするため，より一般的な環境を想定したバンディット問題によって，選択肢の報酬確率分布の平均と分散に差がある場合における解析を行った．その結果，DLR-Q は報酬確率分布の分散に価値の過大評価・過小評価の傾向が比例し，それによって選択肢の弁別性の向上につながっていることを見出した．DLR-Q は複数の報酬確率分布の価値の差を識別するアルゴリズムを検討するための基礎を提供するものと考えられる．また，分散が異なる複数の報酬分布に対し適切な x を設定することで報酬の弁別性を向上させる性質は，報酬の弁別が困難であるより難しいタスクを対象とする深層強化学習などでの応用が期待される．

参考文献

[1] Michael J Frank, Lauren C Seeberger, and Randall C O'reilly. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science (New York, N.Y.)*, Vol. 306, No. 5703, pp. 1940–1943, December 2004.

[2] Romain Cazé and Matthijs van der Meer. Adaptive properties of differential learning rates for positive and negative outcomes. *Biological cybernetics*, Vol. 107, , 10 2013.

[3] Samuel Gershman. Do learning rates adapt to the distribution of rewards? *Psychonomic bulletin & review*, Vol. 22, , 01 2015.

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

[5] Michael J. Frank, Ahmed A. Moustafa, Heather M. Haughey, Tim Curran, and Kent E. Hutchison. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, Vol. 104, No. 41, pp. 16311–16316, 2007.

[6] Michael J Frank, Bradley B Doll, Jen Oas-Terpstra, and Francisco Moreno. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature neuroscience*, Vol. 12, No. 8, pp. 1062–1068, August 2009.

[7] Melissa Bateson. Recent advances in our understanding of risk-sensitive foraging preferences. *Proceedings of the Nutrition Society*, Vol. 61, No. 4, p. 509–516, 2002.