2P-09

報酬量の分散に対する学習率の非対称化による適応

Adaptation by asymmetric learning rate to variance in reward amount

有村 柊一 † Shuichi Arimura　佐鳥 玖仁朗 † Kuniaki Satori　神谷 匠 † Takumi Kamiya　吉田 豊 † Yutaka Yoshida　高橋 達二 † Tatsuji Takahashi　太田 宏之 ‡ Hiroyuki Ohta

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

1. 序論

出産前や飢餓状態などのように, 生物は平常時の意思決定では目標を達成できず, リスクを伴ってでも多くの栄養を必要とする局面に遭遇する [1]. そのような過酷な状態に対応するため, 生物は良い経験と悪い経験を非対称に学習することで適応している可能性が示唆されている [2].

生物の非対称な学習を説明するモデルの一つに, Dual Learning Rate Q(DLR-Q) がある [2][3][4]. 通常の Q 学習は単一の学習率を用いて, 報酬予測誤差の正負に限らず対称に価値を更新する [5]. 一方, DLR-Q では報酬予測誤差の正負によって異なる学習率を用いて非対称に更新する. DLR-Q は, バンディット環境において, 観測した報酬の分散によって価値の過大評価, 過小評価を行い, 選択肢の弁別性を向上させることが確かめられている [6]. しかし, 先行研究ではバンディット環境のみで過大評価, 過小評価を行うことが確かめられており, 状態遷移を伴う強化学習タスクにおいて DLR-Q がどのように振る舞うかは分析されていなかった.

今回我々は, バンディット環境のみで分析されていた DLR-Q を, より生物の生存環境に近いグリッドワールド環境で分析することによって, 生物の目標達成に対して学習の非対称性が及ぼす影響を解析した.

2. 非対称な学習を行う強化学習

ヒトや動物は, 経験から行動を学ぶ. 良い経験に繋がった行動は強化され, 悪い経験に繋がった行動は抑制されることで適切な判断ができるようになる. そのような学習を Q 学習モデルで説明することができ, 良い経験も悪い経験も同一の学習率を用いて学習している. しかし, ヒトや動物は必ずしも良い経験と悪い経験を等しく学習しているわけではないと考えられる. 例えばマウスでは, 良い経験を悪い経験よりも大きく学習することが確かめられている [7]. DLR-Q はそのような, 生物の行う非対称な学習を説明するモデルである.

本研究において DLR-Q は以下のように定義される. ある状態 s_t で行動 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 が以下の式 (1) のように更新される.

$$\begin{aligned} Q_{t+1}(s_t, a_t) =& Q_t(s_t, a_t) \\ &+ \begin{cases} \alpha^+ \Delta Q_t & \text{if } \Delta Q_t \geq 0 \\ \alpha^- \Delta Q_t & \text{if } \Delta Q_t < 0 \end{cases} \\ \Delta Q_t =& r_t + \gamma \max_{a_{t+1}} Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \end{aligned} \tag{1}$$

α^+ と α^- はそれぞれ正の学習率と負の学習率を表し, TD 誤差 (Temporal Difference Error)ΔQ_t が正である際には α^+ で更新され, 負である場合には α^- で更新される. 報酬確率分布に正規分布を用いたバンディット問題で解析した先行研究 [6] によると, DLR-Q は正と負の学習率の比率 $x(= \alpha^+/\alpha^-)$ と報酬確率分布の分散によって価値の過大評価, 過小評価を引き起こす. 価値の過大評価, 過小評価の大きさは分散に比例し, それによって選択肢の弁別性を向上させる. この比例関係は $x < 1$ の時に右肩下がりの傾向となり, $x > 1$ の時に右肩上がりの傾向となる.

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

3. 実験

グリッドワールド環境における DLR-Q の振る舞いを分析しやすいように, 図 1 に示すような簡単な 5×7 のグリッドワールド環境で分析する. グリッドワールドタスクとはエージェントがスタート状態からゴール状態へ向かうことを目的としたタスクで, ゴール状態に到達すると報酬がもらうことができ, エージェントは 1 ステップに 1 マス進むことができる. 実験は, DLR-Q の持つ報酬確率分布の分散によって選択肢の弁別性を向上させるという性質が, 状態遷移の伴う環境で見られるかどうかを確認するため, スタートからの距離と報酬の平均が同じで分散のみが異なる 2 つのゴールで行う. スタート地点からはじまり, 行動によって遷移を繰り返し, いずれかのゴールに到達するとスタート地点へと遷移する. これを 1 エピソードとする.

分析方法として, 正負の学習率比 x について, Q 学習と等価の $x = 1$ を中心とした複数の x で実験することにより, その性質を比較した. 行動空間 A は全ての状態で $A = \{$ 上, 下, 左, 右 $\}$ の 4 つとし, 世界の外に出るような行動は失敗となり, 同じ状態に遷移する. 方策 π は式 (2) に示すように softmax を用いて, 逆温度パラメータ $\beta = 3.0$ とした.

$$\pi(a|s) = \frac{\exp(Q(s,a)\beta)}{\sum_{a \in A} \exp(Q(s,a)\beta)} \tag{2}$$

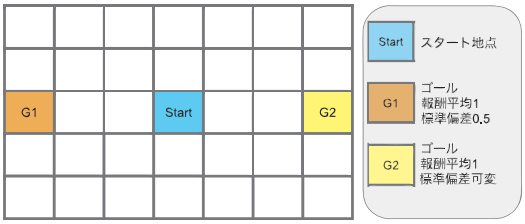

図 1: 実験に使用したグリッドワールド環境

3.1 ゴールの分散のみが異なる実験

G1 のゴールの報酬平均を 1, 標準偏差 σ を 0.5 で固定し, G2 のゴールの報酬平均を 1, σ を 0 から 1.0 まで 0.1 ずつ変化させ, 各 σ で結果の比較を行う. 実験は 10,000 エピソードを 100,000 シミュレーション行った.

図 2 にゴール到達割合のシミュレーション平均の結果を, 図 3 にシミュレーションごとの総報酬獲得量の分布を示す. 図 2 から $x = 1$, つまり通常の Q 学習の時は分散によらずゴールの到達割合が一定であることがわかる. 対して $x > 1$ の時は報酬確率分布の σ が大きいゴールに到達することが多く, $x < 1$ の時は報酬確率分布の分散が小さいゴールへの到達することが多くなっている. これはバンディット環境における性質と一致しており, 状態遷移を伴う環境でも DLR-Q の性質が発揮されていることがわかる.

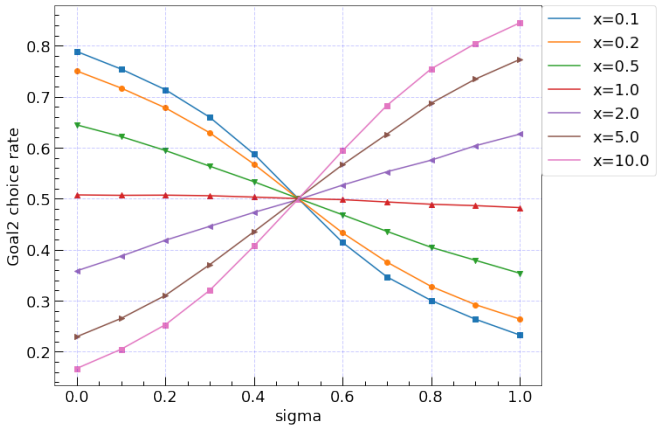

図 2: 各 x ごとのゴール到達割合

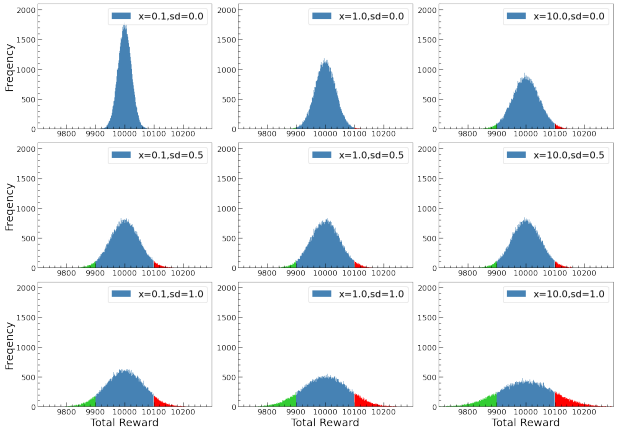

図 3: 各 x, σ ごとの総報酬獲得量の分布

図 3 から x が 10 のときの総報酬獲得量の分布は, いずれの σ であっても, x が 0.1 や 1.0 のときと比べて分布の裾が長くなっている. これは図 2 の到達割合の結果から, x が大きい時に, 分散の大きいゴールへの到達割合が高いため, 総報酬獲得量の分散も大きくなっている.

4. 考察

図 2 から, DLR-Q は, 報酬平均が同じで分散が異なる環境において x が大きいと分散が大きいゴールを選択し, x が小さいと分散が小さなゴールを選択する傾向がある. この傾向はバンディット環境による先行研究 [6] でも見られたものであり, グリッドワールド環境に拡張してもこの性質が失われないことがわかった. この性質は x が小さい時に分散の小さいリスク回避的な選択をし, x が大きいときに分散の大きいリスク選考的な選択をすることを示している.

図 3 から, x が大きいときの総報酬獲得量の分布は x が小さい時の総報酬獲得量の分布と比べて裾が広くなっている. これは x が大きいと分散が大きいゴールへ多く到達するためだと考えられる. またこの結果は, x を大きくした時に, 大きな総報酬となる確率も小さな総報酬となる確率も高めることを示している. これを生物の意思決定と対応づけて, 出産などのために多くの栄養を蓄えるという目標を達成できる総報酬獲得量を 10100 以上, 報酬が少なすぎて生存が困難となってしまう総報酬獲得量を 9900 以下とする. これは図 3 の赤く塗られた部分と緑に塗られた部分にそれぞれ対応しており, x が大きいと, 赤い部分の面積も緑の部分の面積も大きくなっていることがわかる. この傾向は出産前や飢餓状態などのリスクを負ってでも多くの栄養を必要とする生物の意思決定傾向と一致しているため, 良い経験と悪い経験を非対称に学習する DLR-Q というモデルは, 生物の学習をうまく表現できている可能性がある.

5. 結論

本研究では非対称な学習率をもつ DLR-Q の生物の生存環境における振る舞いを明らかにするため, 状態遷移を伴い報酬に正規分布を用いたグリッドワールド環境によって解析を行なった. 結果から, 報酬確率分布の分散と価値の過大評価, 過小評価には比例傾向があり, 選択肢の弁別性が向上するという性質はグリッドワールド環境でも損なわれることなく発揮されることがわかった. x という単一のパラメータによってリスク選好とリスク回避を使い分けられる DLR-Q というモデルは, リスク選好とリスク回避を使い分けなければならないようなタスクで力を発揮することが期待される. また DLR-Q は, 生物の意思決定傾向を表現するモデルとして有効であると考えられるため, 生物の意思決定傾向をモデル化する際の基礎を提供することが期待される.

参考文献

[1] Bateson, M. : Recent advances in our understanding of risk-sensitive foraging preferences, *Proceedings of the Nutrition Society*, Vol. 61, No. 4, pp. 509–516(2002) .

[2] Gershman, J. S. : Do learning rates adapt to the distribution of rewards? , *Psychonomic bulletin & review*, Vol. 22, pp. 1320-1327(2015).

[3] Frank, J. M. , Seeberger, C. L. , and O'Reilly, R. C. : By carrot or by stick: cognitive reinforcement learning in parkinsonism, *Science (New York, N.Y.)*, Vol. 306, No. 5703, pp. 1940–1943(2004).

[4] Caze, D. R. , Meer, M. : Adaptive properties of differential learning rates for positive and negative outcomes, *Biological cybernetics*, Vol. 107, pp. 711-719(2013).

[5] Sutton, S. R. , Barto, G. A. : *Reinforcement Learning: An Introduction.* A Bradford Book(2018).

[6] 佐鳥玖仁朗, 吉田豊, 神谷匠ほか. : 非対称学習率による報酬確率分布の弁別性向上. 情報処理学会第 83 回全国大会講演論文集, 2Q-09(2021)

[7] Ohta, H. Satori, K. , Takarada, Y. , et al. : The asymmetric learning rates of murine exploratory behavior in sparse reward environments. *Neural Networks*, vol. 143, pp. 218-229(2021).