F-006

不完全知覚を含む環境における Q-learning の学習係数に関する一考察

A Case Study on The Learning Rate of Q-learning in Environments involving Perceptual Aliasing

澁谷 長史 †　濱上 知樹 †
Takeshi SHIBUYA　Tomoki HAMAGAMI

1 はじめに

自ら行動し経験を重ねることで振る舞いを獲得する枠組みとして強化学習 [1] が知られている．この枠組みでは，エージェントとよばれる学習主体はある環境のなかで観測・行動・状態遷移を繰り返し（試行錯誤），望ましい状態になった場合には特別な信号（報酬）を受け取る．エージェントはなるべく多くの報酬が得られるような振る舞いの獲得を目指す．

よく知られた強化学習アルゴリズムの 1 つに Q-learning がある．Q-learning では観測と行動の組に対して行動価値を考える．行動価値は将来得られる割引報酬の和を意味する．マルコフ決定過程 (MDPs: Markov Decision Processes) 環境のもとでは Bellman 方程式を解くことによって得られる行動価値を導出できる．この結果から，得られる行動価値が学習係数に依存しないことがわかっている．このことは学習係数の設定を容易にする．

一方，部分観測マルコフ決定過程 (POMDPs: Partially Observable MDPs) 環境のもとでの行動価値の導出方法はこれまで明らかになっておらず，行動価値の学習係数に対する依存性も明らかになっていない．POMDPs 環境において獲得される行動価値を考えることは，POMDPs 環境における学習を実現するうえで重要である．特に筆者らの提案する複素強化学習の枠組み [2] では POMDPs 環境において獲得される行動価値に注目した学習手法であるため，今回検討を行った．

本稿では，簡単な状態遷移構造をもつ環境において獲得される行動価値について議論する．簡単のため，以下のことを仮定する．

- エージェントは予め定められた状態遷移軌道上を決定論的に移動する
- エージェントが各状態において取りうる行動は 1 つである

2 Q-learning

Q-learning では，時刻 t における観測 x_t，行動 a_t に対する行動価値 $Q(x_t, a_t)$ を次式を用いて更新する．

$$Q(x_t, a_t) \leftarrow (1-\alpha)Q(x_t, a_t) + \alpha\left(r + \gamma \max_{a \in \mathcal{A}(x_{t+1})} Q(x_{t+1}, a)\right) \tag{1}$$

† 横浜国立大学大学院工学府

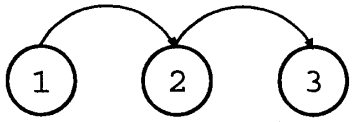

図 1　環境

ここで，r は報酬，$0 \leq \alpha < 1$ は学習係数，$0 \leq \gamma < 1$ は割引率である．

本稿では各状態における選択可能な行動はひとつであり，経路も固定すると仮定しているので，更新式を簡略化できる．観測 x における唯一の行動価値を Q_x と書くと，行動価値の更新は次式で与えられる．

$$Q_x \leftarrow (1-\alpha)Q_x + \alpha(r + \gamma Q_{x'}) \tag{2}$$

ここで，x' は x の次の状態である．

3 MDPs 環境における解析

図 1 のような簡単な環境を考える．本環境において，エージェントはすべての状態を区別できるとする．初期状態は状態 1，終端状態は状態 3 である．状態 3 へと遷移した場合に限りエージェントは報酬 r を獲得する．

行動価値を並べたベクトルを $\mathbf{x} = [Q_1, Q_2, r/\gamma]^{\mathrm{T}}$ とする．ベクトル x に対して，状態 1 から状態 2 へ遷移したときの行動価値の更新（更新 1），状態 2 から状態 3 へ遷移したときの行動価値の更新（更新 2）の 2 つの更新を考える．

更新 1 および更新 2 の更新式は，式 3 および 4 で与えられる．

$$Q_1 \leftarrow (1-\alpha)Q_1 + \alpha\gamma Q_2 \tag{3}$$

$$Q_2 \leftarrow (1-\alpha)Q_2 + \alpha r \tag{4}$$

これらを行列として表現すると，式 5 および 6 となる．

$$\mathbf{x} \leftarrow \begin{bmatrix} 1-\alpha & \alpha\gamma & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x} = T_1\mathbf{x} \tag{5}$$

$$\mathbf{x} \leftarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-\alpha & \alpha\gamma \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x} = T_2\mathbf{x} \tag{6}$$

1 エピソードの更新は，T_1 と T_2 の積で与えられる．$T = T_2T_1$ とすると，1 エピソードを通しての価値の更

新は次式で与えられる．

$$\mathbf{x} \leftarrow T\mathbf{x} \tag{7}$$

T の具体的な形は，次のとおりである．

$$T = T_2T_1 = \begin{bmatrix} 1-\alpha & \alpha\gamma & 0 \\ 0 & 1-\alpha & \alpha\gamma \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

T の固有値は，$\lambda_0 = 1, \lambda_1 = \lambda_2 = 1-\alpha$ である．それぞれの固有ベクトルを $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ とする．これらは $\mathbf{x}$ の空間の直交基底をなすから，任意の $\mathbf{x}$ に対して定数 c_0, c_1, c_2 が存在し，

$$\mathbf{x} = c_0\mathbf{x}_0 + c_1\mathbf{x}_1 + c_2\mathbf{x}_2 \tag{9}$$

とできる．よって

$$T^n\mathbf{x} = c_0\lambda_0^n\mathbf{x}_0 + c_1\lambda_1^n\mathbf{x}_1 + c_2\lambda_2^n\mathbf{x}_2 \tag{10}$$

を得る．いま，$|\lambda_1| = |\lambda_2| < 1$ であるから，$n \to \infty$ においては第 1 項のみが残る．

なお $\mathbf{x}_0$ から，Q_1 および Q_2 を次のように得る．

$$Q_1 = \gamma^2 r \tag{11}$$

$$Q_2 = \gamma r \tag{12}$$

これらは Bellman 方程式から導かれるよく知られた結果と等しい．

4 POMDPs 環境における解析

図 1 の環境において，で状態 1 と状態 2 に不完全知覚が生ずる場合を考える．$\mathbf{x} = [Q_1, r/\gamma]^{\mathrm{T}}$ とする．

ここで更新式に不完全知覚の影響を反映するために，3 つの行列 $\Omega^*, \Omega_1, \Omega_2$ を考える．これらの値は具体的な環境によって異なる．本稿ではこれらを下記の通り与えた．

$$\Omega^* = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{13}$$

$$\Omega_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{14}$$

$$\Omega_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{15}$$

更新 1 および更新 2 において用いる価値の更新に不完全知覚を反映する．このために T_1' および T_2' を次の通り定義する．

$$T_1' = \Omega_1 T_1 \Omega^* = \begin{bmatrix} 1-\alpha+\alpha\gamma & 0 \\ 0 & 1 \end{bmatrix} \tag{16}$$

$$T_2' = \Omega_2 T_2 \Omega^* = \begin{bmatrix} 1-\alpha & \alpha\gamma \\ 0 & 1 \end{bmatrix} \tag{17}$$

$$T' = T_2'T_1' \tag{18}$$

$$= \begin{bmatrix} (1-\alpha)(1-\alpha+\alpha\gamma) & \alpha\gamma \\ 0 & 1 \end{bmatrix} \tag{19}$$

表 1　実験パラメータ

α	γ	r
0.1-0.9	0.2,0.5 0.9	100

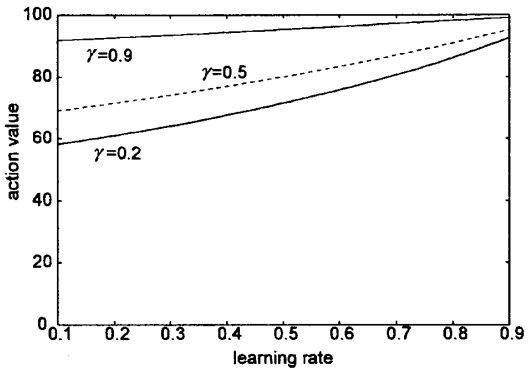

図 2　学習係数と行動価値の関係

T' の固有値は，$\lambda_0 = 1, \lambda_1 = (1-\alpha)(1-\alpha+\alpha\gamma)$ である．$|\lambda_1| < 1$.

λ_0 に対応する固有値 $\mathbf{x}_0$ より，収束値を得る．

$$Q_1 = \begin{bmatrix} \frac{-1}{1-\alpha+\frac{\alpha-2}{\gamma}} \frac{r}{\gamma} \end{bmatrix} \tag{20}$$

5 数値例

結果を図 2 に数値例を与える．用いたパラメータを表 1 に示す．

6 おわりに

簡単な状態遷移グラフをもつ環境において，行動価値を導き，行動価値の学習係数に対する依存性を明らかにした．

MDPs 環境において学習係数がある条件を満足するならば，獲得される行動価値は学習係数に依存しなかった．

POMDPs 環境において獲得される行動価値が学習係数に依存することを主張している．ただし，この環境において $\gamma = 1$ とすれば学習係数に依存しなくなることも明らかとなった．

同じ手法を用いて 4 状態 3 観測の環境における理論値を導き，その後本手法の一般化を行っていく予定である．

参考文献

[1] Richard S. Sutton and Andrew G. Barto, “REINFORCEMENT LEARNING: An Introduction,” MIT Press, 1998.

[2] 澁谷長史，濱上知樹，“複素数で表現された行動価値を用いる Q-learning”，電子情報通信学会論文誌 D, Vol.J91-D, No.5, pp.1286-1295, 2008.