F-033

囚人のジレンマゲームにおけるQ学習による協調の維持

How Does Q-learning Maintain Cooperation in the Prisoner's Dilemma Game?

森山 甲一[†]
Koichi Moriyama

1. はじめに

本稿では，複数の行動主体（エージェント）が，報酬関数も状態変化モデルも未知の状態から強化学習（特にQ学習）を行い，同一の環境中で相互作用を行うマルチエージェント環境を扱う．そのようなマルチエージェント環境におけるQ学習アルゴリズムとして，minimax-Q [3]，Nash-Q [2]，FFQ [4] など，これまでに数多くのものが提案されている．これらの多くは，環境の各状態をゲーム理論における非協力同時手番ゲームと見なし，各エージェントの戦略をナッシュ均衡（合理的プレイヤーの戦略組合せ）に収束させることを目的とする．ところが，囚人のジレンマゲーム [1, 6] のように，ナッシュ均衡が好ましくないゲームというものが存在する．囚人のジレンマゲームは2人2行動ゲームであり，各プレイヤーは「協調」と「裏切り」という2行動を持つ．両者が「協調」することが互いに望ましいが，相手の行動に係わらず「裏切り」を選択する方が利得が大きいため，両者とも「裏切り」を選択することがナッシュ均衡となる．

Q学習は，エージェントがQ関数を元に行動を選択し，得られた報酬からQ関数を更新することで学習する．従って，通常のQ学習エージェント同士で囚人のジレンマゲームを行う際にも,「裏切り」の応酬になることが予想される．しかし，Q学習を行う際には，局所最適に陥らないようにするために行動選択を確率的に行うことが多い．従って，通常のQ学習エージェント同士でも,（ごく小さい）確率で両者が「協調」を選択することがあり得る．そのような協調関係は単発に終わることが予想されるが，仮に，それにより「協調」のQ関数が「裏切り」のものに対して大きくなれば,「協調」が選択されやすくなり，結果として協調関係が維持されるだろう．すなわち，Q学習エージェントによる囚人のジレンマゲームにおける恒常的な協調関係の実現には，1回の行動および報酬がQ関数の更新に及ぼす影響（学習率）が大きな鍵を握っている．

そこで，本稿では，Q学習エージェントが囚人のジレンマゲームにおいて協調関係を維持するための条件について考察する．具体的には，学習率が所与の時に,「協調」のQ関数が「裏切り」のものよりも大きくなる協調関係の回数についての定理を導出する．さらに，筆者はこれまで，エージェントが内部で独自の自己評価（以後効用と呼ぶ）を生成して用いる強化学習により，繰り返し囚人のジレンマゲームではナッシュ均衡に反して「協調」を，それ以外のゲームではナッシュ均衡に相当する行動を学習するエージェントを提案して来た [5]．その観点から，1回の両者の「協調」で協調関係が維持可能となるための効用生成法を提示する．

本稿は4節から成る．2節では準備として囚人のジレンマゲームおよびQ学習を説明する．3節で囚人のジレンマゲームにおけるQ学習エージェントが協調関係を維持する条件について考察し，4節でまとめと今後の課題を述べる．

[†]大阪大学 産業科学研究所

2. 準備

2.1. 囚人のジレンマゲーム

囚人のジレンマゲーム [1, 6] は，2人2行動ゲームの1種であり，**表1**のような利得表と呼ばれる表形式で表記されることが多い．2人のプレイヤー A と B はそれぞれ C と D と呼ばれる2種の行動を持ち，プレイヤー A は行から，B は列から行動を同時に選択する．両者はその行動の組合せから**表1**の利得 $T, R, P, S \in \mathbb{R}$ を得る．但し左側がプレイヤー A，右側がプレイヤー B の利得である．例えば，A が D，B が C を選択した場合，A は利得 T を，B は利得 S を獲得する．以下では X をプレイヤー A の行動，Y をプレイヤー B の行動とした時に，行動組合せを (X, Y) で表現する．

囚人のジレンマゲームとは，利得 T, R, P, S 間に以下の関係を持つゲームのことである．

$$T > R > P > S \quad \text{and} \quad 2R > T + S.$$

この条件により，両プレイヤーはどちらも相手の行動にかかわらず行動 D を選択する方が利得が大きくなる．すると，結果として現れる行動組合せは (D, D) となり，両者とも P の利得を得る．しかし，仮に (C, C) が実現できたとしたら，両者はそれぞれ R の利得を得ることとなるため，こちらの方が望ましい．このように，個々が合理的に行動することで結果として良くない結果をもたらすため，このゲームはジレンマと呼ばれる．

2.2. Q学習

Q学習 [8] におけるエージェントは，離散時間 t における現在の状態 $s_t \in \mathcal{S}$ を知覚し，行動 $a_t \in \mathcal{A}(s_t)$ を選択

表 1　囚人のジレンマゲーム

$A \setminus B$	C	D
C	R, R	S, T
D	T, S	P, P

する．$\mathcal{S}$ は環境の可能な状態の集合を，$\mathcal{A}(s_t)$ は状態 s_t における可能な行動の集合を表す．行動選択後に，エージェントは報酬 $r_{t+1} \in \mathbb{R}$ を受け取り，新しい状態 s_{t+1} を知覚する．これらの情報を元に，Q 学習は行動価値関数 $Q_t(s,a)$ を (1) 式により更新する．状態 s における行動 a の方策 π の下での行動価値とは，行動 a の後に方策 π に従った時に，割引率 $0 < \gamma < 1$ により割り引かれた将来の報酬の和の期待値，すなわち $E\left(\sum_{k=0}^{\infty} \gamma^k r_{t+1+k}\right)$ のことをいう．

$$Q_{t+1}(s,a) = \begin{cases} Q_t(s_t,a_t) + \alpha\,\delta_t & \text{if } (s,a) = (s_t,a_t), \\ Q_t(s,a) & \text{otherwise.} \end{cases} \quad (1)$$

上記で $0 < \alpha \leq 1$ は学習率と呼ばれるパラメータであり，δ_t は $Q_t(s_t,a_t)$ と最適方策における (s_t,a_t) の真の価値の差（TD 誤差）である．

$$\delta_t \triangleq r_{t+1} + \gamma \max_{a \in \mathcal{A}(s_{t+1})} Q_t(s_{t+1},a) - Q_t(s_t,a_t). \quad (2)$$

エージェントが全ての状態への訪問と全ての行動の選択を無限回繰り返し，学習率 α を適切に減少させることにより，環境がマルコフ性を持つならば，全ての s と a について $Q_t(s,a)$ が最適方策における行動価値に収束することが証明されている [8].

最適方策における真の行動価値関数 Q^* が既知ならば，状態 s における最適方策として行動 a^* を以下のように Q^* から導くことが出来る．

$$a^* = \arg\max_{a' \in \mathcal{A}(s)} Q^*(s,a'). \quad (3)$$

ところが，もしエージェントが学習中にこのように選択を行うならば，エージェントが訪れない状態や選択しない行動が現れるかもしれないため，Q_t が局所最適となる可能性がある．これを避けるために，通常はソフトマックス法 [7] のような確率的手法を用いて行動選択を行う．ソフトマックス法は，行動 $a \in \mathcal{A}(s_t)$ の選択確率 $p_a^{s_t}$ を

$$p_a^{s_t} \triangleq \frac{\exp(Q_t(s_t,a)/T)}{\sum_{a' \in \mathcal{A}(s_t)} \exp(Q_t(s_t,a')/T)} \quad (4)$$

により決定する．ここで，$T > 0$ は温度と呼ばれるパラメータで，行動選択のランダム性を制御する．

3. 囚人のジレンマゲームにおける Q 学習による協調の維持

以下では，利得表を知らない 2 台の Q 学習エージェントが，囚人のジレンマゲームを繰り返し行う場合を考える．この時，協調関係 (C,C) を続けさせるためには，

- 協調関係を起こさせ，
- 協調関係を維持させる

ことが必要である．本稿では，前者の実現は偶然に委ねることとし，後者，すなわち Q 学習エージェントが協調関係を維持するための条件について考察する．

以下では**表 1** のプレイヤー A について考える．**表 1** は対称なため，プレイヤー B についても同様に成り立つ．また，行動組合せ (X,Y) におけるプレイヤー A の利得を r_{xy} と表記する．すなわち，$r_{cc} \equiv R$，$r_{cd} \equiv S$，$r_{dc} \equiv T$，$r_{dd} \equiv P$ である．本稿で扱う Q 学習では状態変数を考慮しない．従って，以下では関数 Q の引数は行動のみとなる．学習率 α，割引率 γ をいずれも定数とし，$0 < \alpha < 1$，$0 < \gamma < 1$ とする．

まず，行動価値関数 Q の基本的な性質を 2 つの補題で示す．

補題 1　行動 $X \in \{C,D\}$ の初期 Q 値を $Q_0(X)$ とし，行動 X により得られる利得の最小値を r_x とする．この時，ある時刻 t までに n 回 X を選択した時の $Q_t(X)$ は以下の式を満たす．

$$Q_t(X) \geq (1-\alpha+\alpha\gamma)^n Q_0(X) + \frac{r_x}{1-\gamma}\left\{1-(1-\alpha+\alpha\gamma)^n\right\}. \quad (5)$$

［証明］　X を n 回目に選択した時刻を $f(n)$ とし，$f(0) = 0$ とすると，

$$\begin{aligned} Q_t(X) &\geq (1-\alpha)Q_{f(n-1)}(X) + \alpha(r_x + \gamma Q_{f(n-1)}(X)) \\ &= (1-\alpha+\alpha\gamma)Q_{f(n-1)}(X) + \alpha r_x \\ &\geq (1-\alpha+\alpha\gamma)^2 Q_{f(n-2)}(X) + \alpha r_x\left\{(1-\alpha+\alpha\gamma)+1\right\} \\ &\geq \ldots \\ &\geq (1-\alpha+\alpha\gamma)^n Q_0(X) + \frac{r_x}{1-\gamma}\left\{1-(1-\alpha+\alpha\gamma)^n\right\} \end{aligned}$$

■

補題 2　時刻 t 以降に (X,Y) が連続すると仮定する．但し $X,Y \in \{C,D\}$ である．$Z \in \{C,D\} - \{X\}$ とする．この時，$Q_{t+k}(X) \geq Q_{t+k}(Z)$ for $k = 0,1,2,\ldots$ ならば，

$$\lim_{k\to\infty} Q_{t+k}(X) = \frac{r_{xy}}{1-\gamma}. \quad (6)$$

$Q_{t+k}(Z) > Q_{t+k}(X)$ for $k = 0,1,2,\ldots$ ならば，

$$\lim_{k\to\infty} Q_{t+k}(X) = r_{xy} + \gamma Q_t(Z). \quad (7)$$

［証明］ 上の条件では，

$$\begin{aligned}Q_{t+k}(X) &= (1-\alpha)Q_{t+k-1}(X) + \alpha(r_{xy} + \gamma Q_{t+k-1}(X)) \\ &= (1-\alpha+\alpha\gamma)Q_{t+k-1}(X) + \alpha r_{xy} \\ &= (1-\alpha+\alpha\gamma)^2 Q_{t+k-2}(X) + \alpha r_{xy}\Big\{(1-\alpha+\alpha\gamma)+1\Big\} \\ &= ... \\ &= (1-\alpha+\alpha\gamma)^k Q_t(X) + \frac{r_{xy}}{1-\gamma}\Big\{1-(1-\alpha+\alpha\gamma)^k\Big\} \\ &\to \frac{r_{xy}}{1-\gamma} \quad (k\to\infty).\end{aligned}$$

$Q_{t+k}(Z) = Q_t(Z)$ for $k = 0, 1, 2, ...$ であることに注意すると，下の条件では，

$$\begin{aligned}Q_{t+k}(X) &= (1-\alpha)Q_{t+k-1}(X) + \alpha(r_{xy} + \gamma Q_t(Z)) \\ &= (1-\alpha)^2 Q_{t+k-2}(X) + \alpha(r_{xy} + \gamma Q_t(Z))\Big\{(1-\alpha)+1\Big\} \\ &= ... \\ &= (1-\alpha)^k Q_t(X) + (1-(1-\alpha)^k)(r_{xy} + \gamma Q_t(Z)) \\ &\to r_{xy} + \gamma Q_t(Z) \quad (k\to\infty).\end{aligned}$$

■

この 2 つの補題を用いると，囚人のジレンマゲームにおいて，行動選択が決定的な場合にも，極端な場合を除いて，行動組合せが (C, D) に収束しないことを示すことが出来る．

定理 1 囚人のジレンマゲームにおいて，行動 D の初期 Q 値を $Q_0(D)$ とする．行動選択が決定的，すなわち (3) 式であったとしても，$(1-\gamma)Q_0(D) \geq r_{dd}$ の時，(C, D) に収束することは無い．$(1-\gamma)Q_0(D) < r_{dd}$ の時，$\eta \equiv (r_{dd} - r_{cd})/(r_{dd} - (1-\gamma)Q_0(D))$ とすると，D を少なくとも $\lfloor \log\eta / \log(1-\alpha+\alpha\gamma) \rfloor + 1$ 回実行していれば，同様に (C, D) に収束することは無い．

［証明］ この条件の下で (C, D) に収束するとする．この時，ある時刻 t 以降 C を連続して選択，すなわち任意の $k > 0$ で $Q_{t+k}(C) > Q_{t+k}(D)$ が成立している．補題 2 より $k \to \infty$ で $Q_{t+k}(C) \to r_{cd}/(1-\gamma)$ となる．

一方，$r_{dc} > r_{dd}$ より，時刻 t までに d 回 D を選択している時の Q 値は，補題 1 より以下を満たす．

$$\begin{aligned}&Q_{t+k}(D) = Q_t(D) \\ &\geq (1-\alpha+\alpha\gamma)^d Q_0(D) + \frac{r_{dd}}{1-\gamma}\Big\{1-(1-\alpha+\alpha\gamma)^d\Big\}.\end{aligned}$$

従って，$k \to \infty$ で，

$$\begin{aligned}&Q_{t+k}(D) - Q_{t+k}(C) \\ &\geq \frac{1}{1-\gamma}\Big\{\Big((1-\gamma)Q_0(D) - r_{dd}\Big)(1-\alpha+\alpha\gamma)^d + r_{dd} - r_{cd}\Big\}\end{aligned}$$

となる．

(i) $(1-\gamma)Q_0(D) - r_{dd} \geq 0$ の時，$1-\gamma > 0$，$1-\alpha+\alpha\gamma > 0$，$r_{dd} > r_{cd}$ より，常に右辺 > 0 が成り立つ．

(ii) $(1-\gamma)Q_0(D) - r_{dd} < 0$ の時，

$$\text{右辺} = \frac{r_{dd} - (1-\gamma)Q_0(D)}{1-\gamma}(\eta - (1-\alpha+\alpha\gamma)^d)$$

となり，$d > \log\eta / \log(1-\alpha+\alpha\gamma)$ より，$\eta > (1-\alpha+\alpha\gamma)^d$ が成り立つ．従って，右辺 > 0 が成り立つ．

(i) (ii) より，$k \to \infty$ で $Q_{t+k}(D) > Q_{t+k}(C)$ となるが，これは仮定に反する． ■

定理 1 より，囚人のジレンマゲームにおいて Q 学習プレイヤー同士の対戦を考える場合，ある特定の場合を除けば，行動選択が Q 値に対して決定的であっても (C, D) または (D, C) に収束することはない．例えば，Axelrod による囚人のジレンマゲームトーナメント [1] で用いられた利得表，$r_{cc} = 3$，$r_{cd} = 0$，$r_{dc} = 5$，$r_{dd} = 1$ で $Q_0(D) = 0$ の場合，$\eta = 1$ となり，D を 1 回以上実行していれば定理を満たす．従って，以下では $Q(D)$ として (D, D) が続いた時のものを扱い，それを $Q(C)$ が上回るためには何回 (C, C) が連続すれば良いかを考える．

定理 2 $Q_t(D) = r_{dd}/(1-\gamma)$，$Q_t(C) = r_{cd} + \gamma r_{dd}/(1-\gamma)$，$Q_t(D) > Q_t(C)$ と仮定する．時刻 $t+1$ から $t+k$ まで (C, C) が連続する場合を考える．この時，$Q_{t+k}(C) \geq Q_{t+k}(D) = Q_t(D)$ となる k の条件は，

$$k \geq \frac{1}{\log(1-\alpha)} \log \frac{r_{cc} - r_{dd}}{r_{cc} - r_{cd}} \tag{8}$$

である．

［証明］ $k > 0$ を $Q_{t+k}(C) \geq Q_{t+k}(D)$ を満たす最小の整数とする．$Q_{t+i}(D) > Q_{t+i}(C)$ for $i = 0, 1, 2, ..., k-1$ と補題 2 の証明から，

$$Q_{t+k}(C) = (1-\alpha)^k Q_t(C) + (1-(1-\alpha)^k)(r_{cc} + \gamma Q_t(D))$$

である．$Q_{t+k}(C) \geq Q_{t+k}(D) = Q_t(D)$ から，

$$(1-\alpha)^k\{Q_t(C) - r_{cc} - \gamma Q_t(D)\} \geq (1-\gamma)Q_t(D) - r_{cc}$$

となる．$Q_t(C)$ と $Q_t(D)$ を代入すると，

$$(1-\alpha)^k(r_{cd} - r_{cc}) \geq r_{dd} - r_{cc} \tag{9}$$

から，

$$k \geq \frac{1}{\log(1-\alpha)} \log \frac{r_{cc} - r_{dd}}{r_{cc} - r_{cd}}$$

が導ける． ■

定理 2 は，$Q_t(D)$ が (D, D) に，$Q_t(C)$ が (C, D) に収束する場合の理想値を持つとした時に，$Q(C)$ が $Q(D)$ を上回るために必要となる (C, C) の連続回数を示している．逆に言うと，ある時に (C, C) が起こったら，それ

を k 回以上連続して繰り返すことにより，$Q(C)$ が $Q(D)$ を上回り，結果として (C,C) の選択確率が上がることが期待される．定理 1 の場合と同様に Axelrod の利得表を適用すると，$\alpha = 0.1$ で $k \geq 4$，$\alpha > 1/3$ で $k < 1$ となる．

さらに，利得と効用（自己評価）の関係という視点からは，以下の系が導かれる．

系　定理 2 と同条件で，時刻 $t+1$ に (C,C) となった時，利得 r_{cc} に r を加算することにより $Q_{t+1}(C) \geq Q_{t+1}(D)$ となる r の条件は，

$$r \geq \frac{r_{dd} - (\alpha r_{cc} + (1-\alpha) r_{cd})}{\alpha} \tag{10}$$

である．

［証明］　(9) 式の $k \leftarrow 1$，$r_{cc} \leftarrow r_{cc} + r$ とした

$$(1-\alpha)(r_{cd} - r_{cc} - r) \geq r_{dd} - r_{cc} - r$$

より導ける．■

同様に Axelrod の利得表を適用すると，$\alpha = 0.1$ で $r \geq 7$，$\alpha > 1/3$ で $r < 0$ となる．

4. まとめ

本稿では，Q 学習エージェントが行う囚人のジレンマゲームにおいて，(C,C) が実現された後に，$Q(C)$ が $Q(D)$ を上回るために必要な (C,C) の連続回数，あるいはそのための効用（自己評価）の設計についての条件を導出した．

今後の課題としては，現在は偶然に頼っている，最初の協調関係を実現するための条件の導出が挙げられる．さらに，今回導出された条件では $Q(C)$ が $Q(D)$ を上回るということしか言えない．行動選択が決定的でない場合には，$Q(C) > Q(D)$ であっても D が選択され得るため，必ずしも以後 (C,C) が連続するとは言えず，そのように偶然 D が選択された後に再び (C,C) を続けることが出来るかどうかといった安定性の議論が残されている．また，$Q_t(C)$ および $Q_t(D)$ がそれぞれ (C,D) や (D,D) に収束した時の値を持つという理想的状況における条件でもあるため，実際にエージェントに定理 2 およびその系を適用し，実験的検証をすることも必要である．

ところで，エージェントに自律性を仮定すると，実際に定理 2 およびその系を適用する際には，エージェントは自分が 2 人 2 行動ゲームに参加しているということの他に，

- 現在のゲームが囚人のジレンマゲームであること，
- 今回自分が選択した行動が C, D のどちらに相当するか，
- 今回獲得した利得が利得表（**表 1**）のどこに当たるか，
- 自分が C を選択した時の行動組合せが (C,C) であるか否か，

を知らなくてはならない．ゲーム理論の研究では通常，利得表がプレイヤーに与えられ，プレイヤーはそこからこれらを知ることが可能なため，この問題点は回避されている．もし，エージェントが相手の取った行動を知覚できるのならば，利得と行動組合せから自分の利得表を再構成することにより，近似的にでも知ることが出来るかもしれない．しかし，$Q(C)$ が $Q(D)$ よりも大きくなる理由は，D のもたらす利得が忘れられていくという点にあるため，利得表を知ることによって，協調関係の維持が困難になることも予想される．従って，エージェントが利得表を再構成することなく，上記の点を知ることが出来るかという大きな問いが残されている．

謝辞

本研究は，科学研究費補助金（課題番号 18700145）の助成を受けたものである．

参考文献

[1] R. Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.（松田 訳. つきあい方の科学. Minerva 21 世紀ライブラリー 45. ミネルヴァ書房, 京都, 1998）.

[2] J. Hu and M. P. Wellman. Nash Q-Learning for General-Sum Stochastic Games. *Journal of Machine Learning Research*, 4:1039–1069, 2003.

[3] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. 11th International Conference on Machine Learning, ML'94*, pp. 157–163, New Brunswick, New Jersey, U.S.A., 1994.

[4] M. L. Littman. Friend-or-Foe Q-Learning in General-Sum Games. In *Proc. 18th International Conference on Machine Learning, ICML-2001*, Williamstown, Masachusetts, U.S.A., 2001.

[5] 森山, 沼尾. 自己評価により学習するエージェント. 合同エージェントワークショップ & シンポジウム 2003 (JAWS2003) 講演論文集, pp. 71–78, 兵庫県東浦町, 2003.

[6] W. Poundstone. *Prisoner's Dilemma*. Doubleday, New York, 1992.（松浦 訳. 囚人のジレンマ. 青土社, 東京, 1995）.

[7] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.（三上, 皆川 訳. 強化学習. 森北出版, 東京, 2000）.

[8] C. J. C. H. Watkins and P. Dayan. Technical Note: Q-learning. *Machine Learning*, 8:279–292, 1992.