F-027

強化学習エージェントの協調をもたらす N 人囚人のジレンマゲーム の利得関数

Payoff Function of N-Person's Prisoner Dilemma Game Bringing Cooperation among Reinforcement Learning Agents

田口 智健† 森山 甲一† 武藤 敦子† 松井 藤五郎‡ 犬塚 信博†
Tomotake Taguchi Koichi Moriyama Atsuko Mutoh Tohgoroh Matsui Nobuhiro Inuzuka

1. はじめに

社会において，個人の合理的な選択が社会全体としての最適な選択と一致せずに葛藤が生じる，社会的ジレンマという問題が存在する．社会的ジレンマ問題をモデル化したものとして，囚人のジレンマゲームが存在する．囚人のジレンマゲームにおいて，合理的に行動するプレイヤーは自分が損をしないような行動を取るため，相互裏切りを選択する傾向にある．一方で，森山ら[1,2]は，囚人のジレンマゲームにおいて，プレイヤーが強化学習を行うエージェントの場合に，相互協調が選択されやすい利得の条件を導出した．

囚人のジレンマゲームは 2 人ゲームであるが，それを一般化した N 人囚人のジレンマゲームが存在する．本研究では，N 人囚人のジレンマゲームの利得関数に着目し，エージェントが協調行動を選択するような利得関数の同定を試みる．

2. 準備

2.1 N 人囚人のジレンマゲーム

囚人のジレンマゲームは 2 人のプレイヤーがそれぞれ協調行動 (C) または裏切り行動 (D) を同時に選択し，その組み合わせによって報酬 R ∈ {T, R, P, S} を得るゲームである．ただし$T > R > P > S$であり，行動の組み合わせから得る利得は表 1 のようになる．表より，相手がどちらの行動を選んだかに関わらず自分は D を選択した方がより大きな利得を得られるため，両者は D を選択することが合理的である．しかし相互協調の利得 R は相互裏切りの利得 P より大きいため，相互協調が望まれる．

A/B	C	D
C	R,R	T,S
D	S,T	P,P

表 1：囚人のジレンマゲームの利得表

N 人囚人のジレンマも同様に，N 人のプレイヤーがそれぞれ C または D を同時に選択し，利得を得るゲームである [3]．あるプレイヤーi に着目した時，i が C を選び，i 以外の$v(v < N)$人のプレイヤーが C を選んだ時の i の利得を f(v) とし，i が D を選び，i 以外の v 人のプレイヤーが C を選んだ時の i の利得を g(v) とする．利得関数 f と g において以下の 3 つが成り立つ．

1. $v \geq 0$ の各値に対して$g(v) > f(v)$
2. $f(N-1) > g(0)$
3. f(v), g(v) はともに v の単調増加関数である

このような利得関数 f，g の一例として以下の式が挙げられる．

$$\begin{cases} f(v) = v \\ g(v) = v + b(0 < b < N-1) \end{cases} \quad (1)$$

2.2 Q 学習

Q 学習[4]は強化学習の一種である．強化学習は，行動に対して環境から与えられた報酬を最大化するようなエージェントの学習手法である．Q 学習では，エージェントは現在の状態における選択可能な行動に対して，Q 値または状態行動価値と呼ばれる値に基づいて行動を選択した後，環境から報酬を受け取り次の状態へと遷移し，Q 値を更新する．これを繰り返し行うことでエージェントが最適な行動を学習することができる．

3. 強化学習エージェントの協調をもたらす N 人囚人のジレンマゲームの利得関数

森山ら[1,2]は，囚人のジレンマゲームのプレイヤーに状態数 1 の Q 学習を行うエージェントを仮定し，その上で 1 回の相互協調で$Q(C) \geq Q(D)$となる利得の条件，すなわち，以降に協調行動を促す条件を導出した．本研究では同様に，N 人囚人のジレンマゲームで協調行動を促すような利得関数の同定を試みる．

本研究では(1)式の N 人囚人のジレンマゲームを扱い，切片 b を変更した場合の行動の変化を観察する．シミュレーションの流れは以下の通りである．

1. b の値を範囲内の値で初期化する．
2. Q 値の初期値をランダムな値にして N 体のエージェントを作成する．
3. 作成したエージェントに N 人囚人のジレンマゲームを行わせ，利得関数 f, g から得られた報酬を元に，エージェントに Q 学習を行わせる。
4. 3. を任意回数繰り返し，学習を進める．
5. 2~4 を b の値を範囲内で変更しながら行う．

4. 実験と考察

4.1 実験環境

実験では，作成したエージェントに繰り返し N 人囚人のジレンマゲームを行わせ、エージェントの Q 値の変化や協調行動を選択したエージェント数の動きを観察する．エージェント数 N = 100 とし，ゲームの繰り返し回数を 1000 回として，繰り返し 100 人囚人のジレンマゲームを行った．Q 学習のパラメータは学習率 0.25，割引率 0.5 とする．また，行動選択には $\epsilon = 0.05$ の e-greedy 法を利用し，Q 値の

†名古屋工業大学　大学院工学研究科　情報工学専攻
Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology
‡中部大学　生命健康科学部　臨床工学科　Department of Clinical Engineering, College of Life and Health Sciences, Chubu University

初期値は N 人囚人のジレンマゲームの利得の最大値$(N-1+b)/(1-\gamma)$に [0,1] の一様乱数を掛けたものとした．

4.2 実験結果と考察

図 1 は N 人囚人のジレンマゲームの利得関数において，b の値と 1000 試行の内に協調行動を選択したエージェント数の平均値の関係を表したものである。b が比較的小さい値の時は C を選択するエージェントも多くみられたが，b が大きくなるに連れて D を選択するエージェントが大多数を占める結果となった．

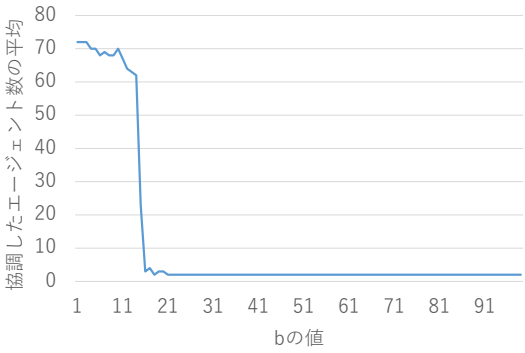

図 1：b の値と協調行動を選択したエージェント数の平均値の関係

図 2 は協調優位の状態から裏切り優位の状態へと転じる b = 14,15の時の協調行動を選択したエージェント数を示したものである．協調優位の状態であるb = 14の時は，一時的に協調するエージェント数が減少することがあるが，再び協調優位の状態へと戻る．しかし，b = 15の時は，次第に協調するエージェント数が減少していき，裏切り優位の状態で収束する．

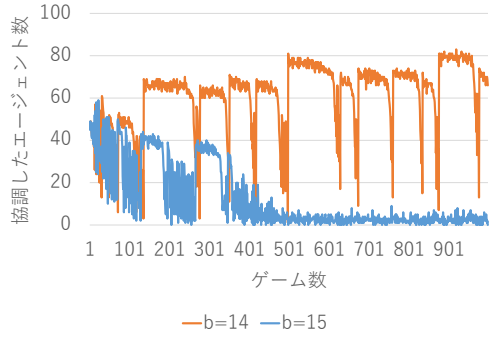

図 2：b = 14,15 の時の協調行動を選択したエージェント数

また，図 3，4 はb = 14,15の時のある一体のエージェント A の Q 値を示したものである．協調優位の状態である b = 14 の時（図 3），最初は初期化された Q 値に基づいてエージェント A は C を取り続けるが，他のエージェントが D を選択するため，次第にQ(C)が低下し，Q(D)がQ(C)を上回った時にエージェント A は D を選択する．当初は協力的な他のエージェントでも同様のことが起き，一時的に一斉に D を選択するが，利得は自分以外の C を選択したエージェント数に比例して与えられるため，Q(D)が低くなる．Q(D) がQ(C)を下回ったエージェントは次のゲームで C を選択するため，また全体として協調優位の状態へと戻る．一方，裏切り優位の状態である b = 15 の時（図 4）を見ると，最初の数ゲーム間は b = 14 の時同じような動きをするが，次第にQ(C)とQ(D)の差が大きくなり，Q(D) > Q(C) となる値に収束する．

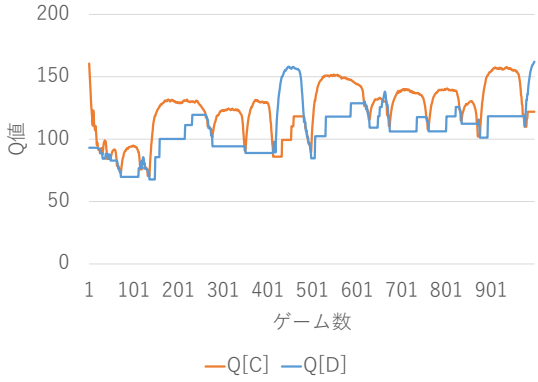

図 3：b = 14 の時のあるエージェント A の Q 値

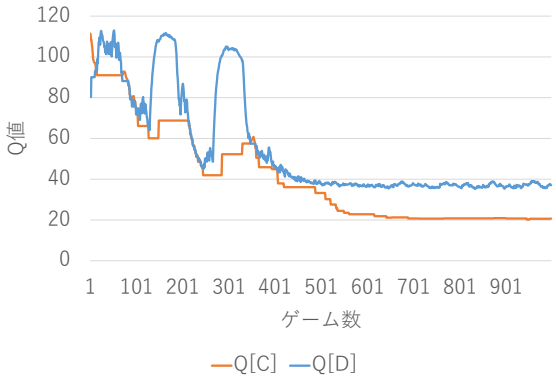

図 4：b = 15 の時のあるエージェント A の Q 値

5. おわりに

本研究では，N 人囚人のジレンマゲームにおいて相互協調をもたらす利得関数の実験的な分析を行った．その結果，協調行動を選択した時の利得と裏切り行動を選択した時の利得の差によって，協調行動を選択するエージェントの数に変化が現れることが分かった．そして，その変化が連続的なものではなく，急激なものであることも確認した．

今後の課題としては，N 人囚人のジレンマゲームにおいて他の利得関数や強化学習手法での分析が挙げられる．

謝辞

本研究の一部は，JSPS 科研費 JP16K00302 の助成を受けて行われた．

参考文献

[1] K. Moriyama: Utility based Q-learning to facilitate cooperation in Prisoner's Dilemma games. *Web Intelligence and Agent Systems*, Vol. 7, No. 3, pp. 233–242, 2009.
[2] K. Moriyama, S. Kurihara, and M. Numao. Cooperation-Eliciting Prisoner's Dilemma Payoffs for Reinforcement Learning Agents. *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1619–1620, 2014.
[3] 松本光崇:N 人囚人のジレンマゲームにおける規範内部化と協調の関係, 人工知能学会論文誌 21 巻 2 号 D, 2006.
[4] C.J.C.H. Watkins and P. Dayan: Technical Note: Q-learning. *Machine Learning*, Vol. 8, pp. 279-292, 1992.