1C3-1

n 人ネットワークの 繰り返し囚人のジレンマゲームにおける利得設計

Reward shaping in the n-person network Iterated Prisoner's Dilemmas

鈴木香名子　荒井幸代
Kanako Suzuki　Sachiyo Arai

千葉大学大学院工学研究科
Graduate School of engineering, Chiba University

This paper takes a reward shaping problem in a context of multiagent's reinforcement learning system which consists of individually-motivated agents. Specifically, we focus on the domain where Q-learning agents are connected by a network, and are affected more or less by their neighborhood, under the n-person Iterated Prisoner's Dilemmas, the difficult game in the sense that agents could not choose an optimal actions. In such a domain, three factors, i.e.; a payoff structure, an agent's decision criterion and a network structure, are generally introduced to consider the agents' interaction. These factors largely affect the behavior of multiagent system. In addition to these factors, we consider the "reward shaping problem" to make system behave cooperative In the previous game theoretic approaches, which introduce Q-learning, take two players cases. However, these approaches seem difficult to apply to larger scale multiagent systems.

In this paper, n agents play the Prisoner's Dilemma with neighbor agents opponents. First, we show two types of Q-values updating, synchronous and non-synchronous. Second, we discuss the influence of the network structure, and reward shaping effects on multiagent's learning.

1. はじめに

マルチエージェントシステムの挙動を向上させることを目的として，エージェントの行動設計に強化学習を適用する試みが増えている。しかし，ゲーム理論で議論されてきた社会的ジレンマは，そのまま，マルチエージェントシステムへの強化学習の適用においても大きな問題である。すなわち，エージェントの行動ルールを設計するために強化学習を用いると，エージェントが自らの利益を最大化しようとして系全体が最適ではない状態に陥るようなジレンマが生じるのは当然である。

そこで，本論文では，強化学習において周囲のエージェントの行動やそれらが得た報酬を，情報（状態入力）として，エージェント間で共有可能なマルチエージェントモデルを考える。具体的には，マルチエージェントシステムにおいて，強化学習このようなジレンマを解消するためには，エージェント間に存在する利害関係，どのエージェントと利害関係にあるかを表すネットワーク構造が各エージェントの学習に与える影響を考察する。

本論文では，エージェント間の利害関係を囚人のジレンマゲームによって定式化し，強化学習の手法である Q 学習を用いてエージェントが系全体が最適になるような行動ルールを獲得することを目指す。エージェント数が 2 の環境を対象とした森山による効用と呼ばれる自己評価を用いた報酬設計法 [森山 07] を基本として，エージェント数が 3 以上の環境での学習の更新のタイミング，報酬の与え方，効用の与え方について実験的に考察する。

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では 2 人のプレイヤー A と B が協調行動 (C; Cooperate) と利己的行動 (D; Defect) と呼ばれる 2 種の行動のいずれかを同時に選択する。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，行動の組み合わせを (x, y) と表す。両者はその行動の組み合わせから表 1 に基づき，それぞれ利得 r_{xy}, r_{yx} $(r_{xy}, r_{yx} \in \{T, R, P, S\})$ を得る。ここで，r_{xy} はプレイヤー A の利得, r_{yx} はプレイヤー B の利得である。

表 1: 囚人のジレンマゲーム 利得行列

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

囚人のジレンマゲームでは利得 T, R, P, S 間に $T > R > P > S$ かつ $2R > T + S$ という関係が成り立つ。この条件の下では両者とも相手の行動に関わらず D が支配戦略であるためナッシュ均衡は (D,D) となるが，パレート最適は (C,C) である。

2.2 Q 学習

強化学習の手法として Q 学習を用いる。エージェントは状態 $s \in \mathcal{S}$ を知覚し，方策 π に基づいて行動 $a \in \mathcal{A}(s)$ を選択する。ただし，$\mathcal{S}$ は環境の遷移可能な状態の集合を，$\mathcal{A}(s)$ は状態 s において選択可能な行動の集合を表す。エージェントは行動選択後に報酬 r を受け取り，新しい状態 s' を知覚する。Q 学習は状態 s-行動 a の価値 $Q(s, a)$ を式 (1) により更新する。ここで $\alpha(0 < \alpha \leq 1)$ は学習率，$\gamma(0 \leq \gamma \leq 1)$ は割引率を表し，k は s において a を選択し，$Q(s, a)$ を更新した回数である。

$$Q_{k+1}(s, a) = Q_k(s, a) + \alpha\{r + \gamma \max_{a' \in \mathcal{A}(s')} Q_k(s', a') - Q_k(s, a)\} \quad (1)$$

連絡先: 鈴木香名子，千葉大学大学院工学研究科，千葉市稲毛区弥生町 1-33，043-251-1111(代表)

3. 問題設定

3.1 対象環境

エージェント数が n (≥ 3) の環境において，各エージェントは特定の相手とそれぞれ囚人のジレンマゲームを繰り返し行う。本研究で用いる囚人のジレンマゲームの利得行列は，表 1 より $R = 3, S = 0, T = 5, P = 1$ とする。また，各エージェントとその対戦相手との関係を図 1 のようなネットワークとして表現する。図 1 において，ノードはエージェント，リンクはゲームを行う相手との接続関係を表す。

エージェント i ($= 1, 2, \cdots, n$) の次数を δ_i と表すと，系全体で行われるゲームの数は $\sum_{i=1}^{n} \delta_i/2$ である。ここで次数とは，各エージェントに接続したリンクの数を示しており，各エージェントが実行するゲームの数である。系で行われるゲームを g_j ($j = 1, 2, \cdots, \sum_{i=1}^{n} \delta_i/2$) と表す。また，エージェント i がゲーム g_j で得た利得を r_{ij} と表す。

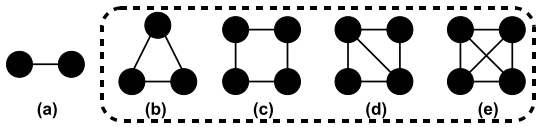


図 1: 対象環境

3.2 関連研究

森山はエージェント数が 2 の環境において Q 学習を用いて PD における協調を維持する手法として，偶然に相互協調 (C,C) が起きた時に，効用と呼ばれる内部報酬を導入して C の価値を上げる方法を用いた。ここでの Q 学習は状態変数を考慮しないため Q 値の引数は行動だけとなり，それぞれ Q(C), Q(D) と表す。

効用とはエージェント内部の刺激であり，エージェントの外部からもたらされる報酬によって生じる。ここでは，効用をゲームの結果獲得する利得とパラメータ r' の和とし，それを Q 学習における報酬として用いる。ある時刻 t において，Q_t(D) $> Q_t$(C) と仮定する。$t+1$ に (C,C) となった時，利得 R 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' の条件は式 (2) となる。

$$r' \geq \frac{P - \{\alpha R + (1-\alpha)S\}}{\alpha} \tag{2}$$

森山はエージェント数が 2 の環境しか扱っておらず，エージェント数が大きな環境への適用可能性には言及していない。

一方，石渕ら [石渕 00] や鈴木ら [鈴木 01] のように空間構造や空間局所性がエージェントの戦略に与える影響を解析した研究がある。石渕らの扱った空間型繰り返し囚人のジレンマゲームでは，複数のプレイヤーが存在する空間を構造化し，各プレイヤーは隣接するプレイヤーとのみ対戦を行う。しかし，これらの研究は遺伝的アルゴリズムによる戦略進化を行った場合の協調行動の創発への影響を解析したものであり，エージェントが自律的に協調行動を学習できるかとは別の枠組みである。

本研究では学習による協調行動獲得とネットワーク構造の影響の立場から，森山の効用を用いた報酬設計法をベースに，エージェントを 3 人以上に拡張した場合の協調維持の可能性を実験的に考察する。

3.3 拡張法

エージェント数が 3 以上の環境において各エージェントが Q 学習を行うが，Q 値の引数は森山の手法にしたがい行動のみとする。学習を行うタイミングを非同期更新による学習と同期更新による学習の 2 通りを設定する。

1. 非同期更新による学習　各エージェントはそれぞれのゲームごとに行動を選択し，報酬を受け取り，Q 値を逐次更新して学習を進める。ゲームを行う順番はランダムとする。それぞれのゲームごとに Q 値を更新し学習を行うため，森山の効用の考え方をそのまま適用できると考える。

2. 同期更新による学習　各エージェントは自身の Q 値にしたがって行動を 1 つ選択し，すべてのゲームを同時に行う。その各ゲームで得た利得 r_ij を合計した値を報酬として Q 値を更新する。このとき，効用を用いるタイミングを 2 通り設定する。

設定 1　系全体が協調した場合に効用を用いる

設定 2　2 エージェント間のゲームにおいて相互協調 (C,C) が起きた場合に効用を用いる

森山は相互協調を起こすのは偶然に委ねていた。本研究も基本的にこれにしたがうが，エージェント数を n 人に拡張した場合，系全体が偶然協調になる確率は指数関数的に減少する。そこで設定 1 を緩和して，2 エージェント間のゲームにおいて相互協調 (C,C) が起きた場合に効用を用いる方法を設定 2 とし，実験によって設定 1 と比較する。

4. 実験 1: 同期更新による学習

4.1 実験設定

同期更新による学習では，各ゲームで得られた利得を合計したものを報酬とするため，森山の効用の考え方をそのまま適用することはできない。そのため，図 1(a) に示した $n = 3$ の環境において，効用のパラメータ r' の値，α の値，効用を用いるタイミングを検証する実験を行う。

系全体で行われるゲームをそれぞれ 20000 回繰り返し，これを 1 試行とする。効用によって協調を維持できるかを評価するため，15001 回目以降のゲームにおいて系全体が協調する割合 P_{C} を評価する。乱数を変えた 500 試行を行い，平均値を結果として用いる。ただし，15000 回目までに効用を用いる機会が無かった試行は除外する。学習率 $\alpha \in (0, 1)$ を 0.05 刻み，効用のパラメータ $r' \in [0, 2)$ を 0.1 刻みで変化させ，$\gamma = 0.5$ と 0.9 の各場合について実験を行う。行動選択には $\varepsilon = 0.05$ の ε-greedy 法を用いる。

4.2 実験結果と考察

それぞれの γ の値に対して，設定 1 によって効用を用いた実験結果を図 2 に，設定 2 によって効用を用いた実験結果を図 3 に示す。図 2, 3 それぞれにおいて，横軸は学習率 α，縦軸は全エージェントが同時に C を選択する割合 P_{C} を表す。また，設定 2 においては 500 試行のうち結果から除外した試行は無かったが，設定 1 においては平均して 398.7 回の試行しか評価の対象となっておらず，残りの 101.3 回は 15001 回までのゲームにおいて効用を用いる機会が無かった。

図 2，3 より，効用の値を大きくすることによって，α の値に依存せずに系全体の協調が維持される割合が増加した。また，設定 2 の $\alpha = 0.05$ の場合を除いて，$\gamma = 0.9$ の方が良い結果を示した。図 2, 3 の比較より，設定 2 の法が設定 1 より系全体の協調が起きる割合が大きい。さらに設定 2 ではほぼ全試行が評価の対象となることから，設定 2 の方が設定 1 よりも優れていると言える。

学習における設定 1 と設定 2 の比較　$\alpha = 0.3$, $r' = 1.9$, $\gamma = 0.9$ を用いた場合，図 2, 3 より，設定 1 では約 46.7%, 設定 2 では約 94.1%の割合で系全体の協調行動が起きている。

設定 1 を用いた場合，全員が Q(D) $> Q$(C) の状態で偶然系全体の協調が起きたとき，3 人が同時に効用を用いることで

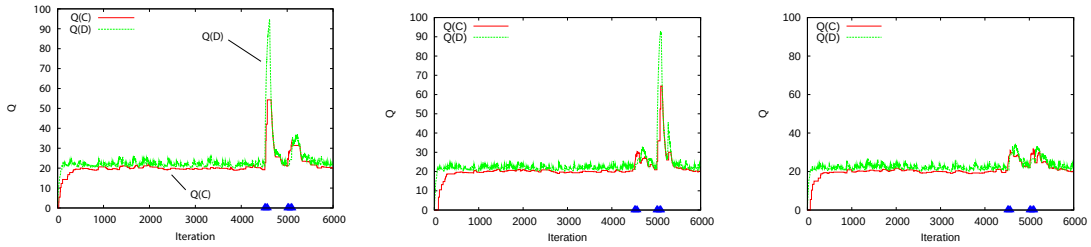
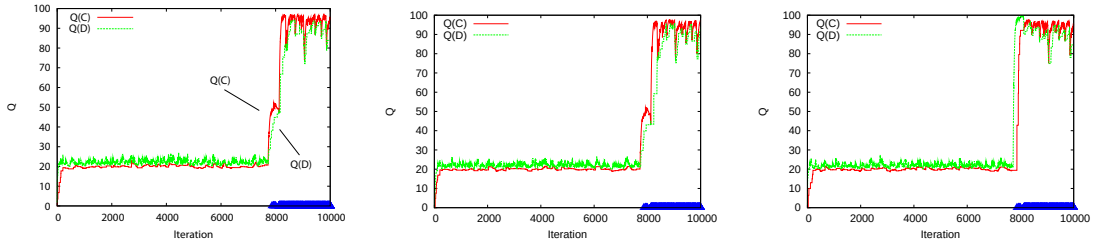
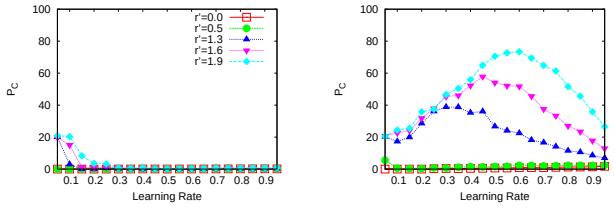


図 4: 設定 1 で協調が維持されなかった試行



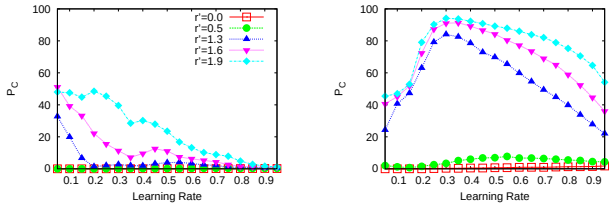
(a) エージェント 1　(b) エージェント 2　(c) エージェント 3

図 5: 設定 2 で協調が維持された試行



(a) $\gamma = 0.5$　(b) $\gamma = 0.9$

図 2: 設定 1 を用いた結果



(a) $\gamma = 0.5$　(b) $\gamma = 0.9$

図 3: 設定 2 を用いた結果

$Q(\mathrm{C}) \geq Q(\mathrm{D})$ となった試行においては，それ以降のゲームで系全体の協調が維持されている。しかし，3 人が同時に効用を用いたとしても全員が $Q(\mathrm{C}) \geq Q(\mathrm{D})$ にならないことがある。もし 1 人のエージェントが効用によって $Q(\mathrm{C}) \geq Q(\mathrm{D})$ とならなかった場合，それ以降のゲームにおいて 2 人のエージェントが C を選択しても 1 人のエージェントが D をとり続け，結果として系全体がパレート劣位に陥るため効用の効果が小さい。この時の 3 エージェントの Q 値の推移を図 4 に示す。図 4 において，横軸はゲームの回数，縦軸は Q 値を表す。また，系全体の協調が起きた回数を x 軸にプロットした。

対して設定 2 は，2 エージェント間のゲームにおいて効用を用いることで 1 人が利己的行動をとってもその 2 人の相互協調は維持され，系全体の協調行動をうながす。この時の 3 エージェントの Q 値の推移を図 5 に示す。

以下では特に設定 2 について考察する。

割引率 γ の影響　図 2, 3 より $\gamma = 0.5$ の場合に比べ，$\gamma = 0.9$ では協調行動を維持しやすいことがわかる。これは効用により $Q(\mathrm{C}) \geq Q(\mathrm{D})$ となった後，γ が大きいほど $Q(\mathrm{C})$ が下がりにくいためである。効用により $Q(\mathrm{C}) \geq Q(D)$ となった後でも確率 ε でランダムな行動を選択するため，行動 D を選択することがある。このとき，γ が大きいほど対戦相手に D を選択されても $Q(\mathrm{C})$ が下がりにくくなり，結果として C を選択する割合が大きくなる。

学習率 α の影響　α の値によって大きく学習の様子が異なる。α が大きいと効用を与えなくても容易に $Q(\mathrm{C})$ と $Q(\mathrm{D})$ の大小関係が入れ替わるため，効用を与えることによる変化は小さい。逆に α が小さいと $Q(\mathrm{C})$ と $Q(\mathrm{D})$ の大小関係は学習の初期段階にとった行動に依存し，結果として系全体が C または D に収束する。D に収束した場合 $Q(\mathrm{C}) \geq Q(\mathrm{D})$ とするには効用のパラメータ $r' \geq 2$ を与える必要があるが，$R + r' \geq T$ となり各エージェントが囚人のジレンマゲームを行っているとは言えなくなる。しかし，実験では $r' \in [0, 2)$ の範囲において r' が大きいほど C に収束する割合が上がることがわかった。この結果より，α が小さい場合においては学習の初期段階にとる行動を制御することによって系全体の協調を実現できる可能性があると考える。

5. 実験 2: ネットワーク構造による影響

5.1 実験設定

図 1(a)〜(e) の環境を用いて，ネットワーク構造の違いを比較する実験を行う。非同期更新による学習と同期更新による学習の 2 つの拡張方法を用いる。以下ではそれぞれを同期，非同期と表記する。

評価方法，行動選択法は実験 1 と同様とし，実験 1 において最も系全体が協調する割合が大きい $\alpha = 0.3$, $\gamma = 0.9$ を用いる。効用のパラメータ $r' \in [0, 2)$ を 0.1 刻みで変化させ実験を行う。

5.2 実験結果と考察

各環境に対する実験結果を図 6, 7 に示す。図 6, 7 において，横軸は効用のパラメータ r'，縦軸は全ゲームで相互協調が発生した割合 P_{C} を表す。500 試行のうち，結果から除外した試行は無かった。

図 6, 7 より，非同期と同期のそれぞれの拡張方法ではどちらもエージェント数が多い環境ほど系全体が協調する割合が減っている。しかし，図 1(c)〜(e) の $n = 4$ の環境において，系全

体が協調する割合に違いがある。同期では各エージェントの次数が多い環境ほど協調する割合が大きいのに対し，非同期では次数が少ない環境ほど協調する割合が大きい。

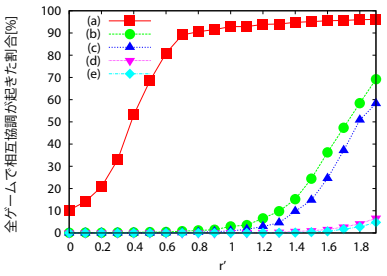


図 6: 非同期更新の結果

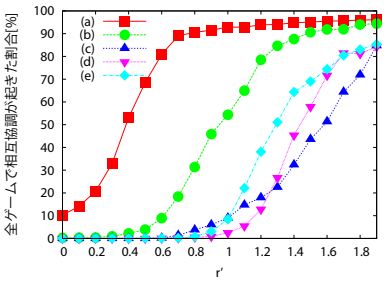


図 7: 同期更新の結果

非同期更新による学習　非同期更新では学習を行うゲームの利得行列が変わらないため，森山の導出した定理を用いることで協調行動を維持できる可能性があると考えていた。式 (2) より，$\alpha = 0.3$ の場合 1 回の相互協調 (C,C) で $Q(\mathrm{C}) \geq Q(\mathrm{D})$ になるには効用のパラメータ $r' \geq 1/3$ を用いる必要がある。図 6, 7 より，図 1(a) の $n = 2$ の環境においては $r' \geq 0.7$ を用いることで 90%以上の割合で全ゲームで相互協調が起きた。$r' \geq 1/3$ である $r' = 0.4$ では約 50%しか系全体の協調が起きていないのは，実際の学習では時間によって Q 値が変動するため，相互協調が起きても $Q(\mathrm{C}) \geq Q(\mathrm{D})$ にならない場合があることが理由である。

また，エージェント数が多い環境ほど系全体の協調が起きる割合が極端に減少する。これは，一方の相手との相互協調が発生し $Q(\mathrm{C}) \geq Q(\mathrm{D})$ になったとしても，他方の相手に D をとられれば $Q(\mathrm{C})$ は下がり，前回相互協調が起きた相手と再び対戦するまでに互いが C を選択する確率が低くなるためである。実験結果では，各エージェントにつながるリンクの数が多い環境ほどこの傾向が見られ，図 1(d), (e) の環境では効用による効果がほとんど見られなかった。

同期更新による学習　同期更新では，$n = 4$ の環境に対して，エージェント同士がつながるリンクが多い環境の方が系全体の協調が維持されやすいことがわかった。図 1(d), (e) ではクラスターを構成する 3 エージェントで協調を維持する様子が見られた。これは誰かが D を選択しても 3 人で協調を維持することができることと，D を選択したエージェントに対して 3 人が影響を持つことが理由である。

6. まとめと今後の課題

本研究の目的は，複数の相互に関係し合うエージェントが存在するシステムを対象とし，エージェント間の協調の実現である。複数エージェントの協調に関する理論的研究としてゲーム理論があるが，エージェント数を 2 とした議論が主流で，大規模なマルチエージェントシステムに対して直接適用することは難しい。

本研究では，現実の環境における協調の実現に向けて，エージェント数が 3 以上の環境において，ジレンマを解消するための報酬設計を森山の効用と呼ばれる内部報酬を導入して実験的に考察した。

また，エージェント数増加に伴って考慮すべきこととして，各エージェントの学習タイミングに着目し，各エージェントが戦略を同期して更新する方法 (同期更新) と，独立に更新する方法（非同期更新）の 2 つを提案した。

同期更新による学習ではそれぞれのゲームごとに Q 値を更新しないため，森山の導出した定理を用いることができないことを示した。一方，非同期更新による学習ではそれぞれのゲームでは森山の定理を用いることができるが，効用を用いて $Q(\mathrm{C}) \geq Q(\mathrm{D})$ になったとしても，エージェント数が 3 以上の環境では，系全体の協調を維持することはできないことを示した。

以上の結果から，森山の考え方をそのままエージェント数が 3 以上の場合に適用することはできなかったが，効用の値や学習率 α を調節することでエージェント数が 3 以上の環境においても系全体の協調が維持されることを確認した。

学習パラメータについて，α が小さい場合のように，囚人のジレンマゲームの条件に反しない範囲で効用を与えるだけでは，系全体の協調を維持できない場合があることを示した。

今後の課題として，学習をパレート最適な状態に収束させるための報酬の与え方やタイミング，行動選択法を含めた報酬設計を考えることを挙げる。また，本研究では，協調を維持させることだけに言及し，学習の収束性については議論をする必要がある。これらの理論的な解析を進めると同時に，今回扱ったエージェント数が 4 よりもさらに多い環境において，ネットワーク構造やエージェント数が各エージェントの学習に与える影響を確かめることが必要である。

参考文献

[森山 07] 森山甲一: 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，合同エージェントワークショップ&シンポジウム 2007 講演論文集，2007.

[Babes 08] M. Babes, E. Munoz, and M. L. Littman: Social Reward Shaping in the Prisoner's Dilemma, Proceedings of the Seventh international joint conference on Autonomous agents and multiagent systems-Volume3, pp.1389-1392, 2008.

[強化学習] R.S.Sutton and A.G.Barto: Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998. (三上貞芳，皆川雅章訳，強化学習，森山出版，東京，2000).

[石渕 00] 石渕久生，中理達生，中島智晴: 空間型繰り返し囚人のジレンマゲームにおける隣接プレーヤー間での信頼関係のモデル化，電子情報通信学会論文誌 D-I, Vol.J83-D-1, No.10, pp.1097-1108, 2000.

[鈴木 01] 鈴木麗璽，有田隆也: N 人版繰り返し囚人のジレンマゲームにおける空間的局所性とその進化，電子情報通信学会技術研究報告，2001.