報酬変動型繰返しゲームにおけるエージェントの協調行動の性質

稲垣裕伸[†] 魚住超[††] 小野功一[††]

この論文は，強化学習を行う2つのエージェントが，動的な環境で活動するシミュレーションについて報告し，それぞれが協調するための条件について議論している．囚人のジレンマなど，これまでのアプローチは，エージェントが置かれた環境が不変であるものと想定していたが，本研究では，エージェントの行動によって，環境が変化するような繰返しゲームの考え方をシミュレーションに導入し，2つのエージェントが協調できるのは，どのような状況かを考察した．

The Agents' Characteristics Which Generate Cooperative Behavior in an Iterated Game with Variable Payoff

HIRONOBU INAGAKI,[†] TAKASHI UOZUMI[††] and KOICH ONO[††]

This paper reports the results of the simulations in which two agents that have ability of reinforcement learning act to achieve their own goal in dynamic environment and discusses about the conditions under which two agents can cooperate with each other. We assumed the dynamic environment with the iterated game whose payoff values are changed by agents action and discussed the situation in which two agents can cooperate.

1. はじめに

自然界では，多くの生物が各々の生命を維持するために競合したり協調したりする様子がみられる．そのため生物種の個体数や環境などはつねに変動しているが，生態系全体の恒常性が維持されている例が多くみられる．これは自然界に限らず，人間社会のなどの複雑なシステムにもみられる．

このような複雑なシステムで，システムの整合性などの深い配慮を考えない場合は，その各構成員は生命を維持するなどの，各々の利己的な目標をかなえるための行動をとっていて，その行動の結果が，巡りめぐって自身や他の構成員に影響を及ぼしていると考えられる．各構成員は，環境や他の構成員に容認され，同時に自身の利益も導くような行動を，長い時間で獲得した結果，全体の恒常性が維持できるシステムができあがったと予測できる．

† 室蘭工業大学博士後期課程生産情報システム工学
Doctor Course Student of Production Information System Engineering, Muroran Institute of Technology
†† 室蘭工業大学情報工学科
Department of Computer Science and System Engineering, Faculty of Engineering, Muroran Institute of Technology

このように，恒常性を維持できる複雑なシステムには，各構成員の行動が間接的に他者に影響を与え，自己にも及ぶという，フィードバック機構の複雑なネットワークがはりめぐらされているという特徴がある．

これは自然のシステムだけではない．たとえば，将来的にはインターネットでもソフトウェアエージェントの社会ができあがるものと考えられているが，そのための社会レベルでのエージェントを真剣に考察すべきだという議論もある[1]．

本研究では，ソフトウェアに限らず，各々の目標を達成するために行動し，その結果状況を遷移させて他の構成員や自分にも影響を及ぼす主体をエージェントとして扱うが，このようなエージェントどうしのかかわりを解析する有効な手法にゲーム理論がある．とりわけ囚人のジレンマゲームは，これまでマルチエージェントの問題としてもゲーム理論の問題としても多数議論されてきた[2)~5)]．

しかし，これまでのゲームは，各エージェントの利得値が変化しないことを前提にしているが，実際にエージェントが活躍する環境は，一定で変動しないことを想定することはできない[2)]．エージェントを取り巻く環境が変化するような問題は，文献2), 3), 6)など，数多く議論されているが，エージェントどうしの

関係が状況に応じて変化するような議論は少ない．

また，変化する環境についての情報をすべて実時間で集められるとは限らない．むしろ自然界の複雑システムは，その構成員が，つねに変化する環境に対して，適した行動をとり続けるよう学習したことによって，恒常性を維持できる仕組みを獲得したものと考えられる．

そこで本研究では，利得の変化する繰返しゲームで動的環境を表現し，その環境において，強化学習を行う複数のエージェントが，それぞれ独自の目標を達成しようとするシミュレーションを行った．そして，強化学習を行うエージェントが，動的な環境の中で協調していくための条件を，各エージェントの目標と性質の観点から考察した．特に，環境や他者に容認されるような寛容な性質を持っているエージェントが存在する場合，全体を安定に保つことができるかどうかを確認した．

2. 基本的な考え方

ある目標を達成しようとするエージェントは，目標についての状態に関心を持っているといえる．たとえば，ソフトウェアエージェントではないが，循環器系における自律神経のように，心拍数を上昇させる活性系と，心拍数を下降させる抑制系が，心拍数をそれぞれの目標値に維持しようとしているような場合を考える．このときエージェントが関心を持っている対象の状態は，時間的に変化するので，エージェントはつねに同じ環境の状態を期待できるわけではない．そのため，ここでは興味対象の状態によって環境の状況がまったく異なるということを示す．

いま，活性エージェントAと抑制エージェントSの2つのエージェントがいて，対象の状態（心拍数）x_tを自分たちの目標値に近づけようとしているとする．抑制エージェントの目標値をg_s，活性エージェントの目標値g_aとする．このとき両者の目標値は不変であるとし，$g_s < g_a$であるとする．簡単のため，両者の行動は「活動しない」(N)と「活動する」(W)の2つであるものとし，抑制エージェントが活動するとx_tを下降させ，活性エージェントが活動するとx_tを上昇させるものとする．ゲームは非協力ゲームで，交渉などは行わないものとする．

x_tが両者の目標値から離れているときと，両者の目標値の間にあるときの利得表を**表1**と**表2**に示す．両エージェントが同時に活動すると，両者の作用が打ち消し合うので状態は変化しない．両者が活動しないときも同様である．どちらか一方のみが活動すると上昇したり下降したりする．その結果，x_tがそれぞれの目標値に近づくと，得をし，遠ざかると損をすると考えることにする．

表1 $x_t > g_a > g_s$のとき
Table 1 $x_t > g_a > g_s$.

(S, A)	N	W
N	(0, 0)	(−1, −1)
W	(1, 1)	(0, 0)

表2 $g_a > x_t > g_s$のとき
Table 2 $g_a > x_t > g_s$.

(S, A)	N	W
N	(0, 0)	(−1, 1)
W	(1, −1)	(0, 0)

表1の状況で，ゲーム理論による解析を行ってみると，エージェントS，Aの各行動はそれぞれ(W, N)で均衡する．これは抑制エージェントだけが活動して活性エージェントが活動しないとき，両者の目標値に近づけることができる行動である．一方，表2の状況では，エージェントS，Aの各行動はそれぞれ(W, W)で均衡する．この状況では，どちらか一方だけが活動すると，活動した方のみが得をし，活動しない方は損をするため，互いに行動Wを選択しなくてはならない．表1の状況の各行動は，両者が各々の目標に近づけようとしているので協調しているように見えるが，表2の状況では，互いに競合しているように見える．

このように，x_tの値によって状況は変化する．

動的な環境では，興味対象の状態は，当事者であるエージェントやその他多くの要因の関与からつねに変化すると考えられる．そのため状況もつねに変化し，最適な行動が一定ではない．そこで動的な環境を表現する方法として，ゲームの利得表の利得値が，各エージェントの行動によって変化するような繰返しゲームを考える．

ここで重要なのが，それぞれのエージェントの利得関数と行動である．つまり，各エージェントがどの程度目標を達成できたときには，どれくらい得したり損をしたりしたと考えたらよいのかという評価の問題のことである．これは，エージェントを適用する問題ごとに設定しなくてはならいと考えられるが，ここでは活性エージェントAの利得関数をf_a，抑制エージェントSの利得関数をf_sとする．これらの関数は，x_tをどの程度各々の目標値に近づけたかを評価する関数であるから，x_tの関数となる．よって，それぞれのエージェントは，各時間で$f_a(x_t)$，$f_s(x_t)$の利得を得ることになる．

各エージェントの行動は「活動しない」(N)か「活

表 3 利得（環境）変動型繰返しゲームの利得行列
Table 3 A payoff matrix of the iterated game with variable payoff (environment).

(S, A)	N	W
N	$(f_s(x_{nn}), f_a(x_{nn}))$	$(f_s(x_{nw}), f_a(x_{nw}))$
W	$(f_s(x_{wn}), f_a(x_{wn}))$	$(f_s(x_{ww}), f_a(x_{ww}))$

動する」(W) の 2 つであるので，現在の状態量 x_t から次の時間の状態への遷移は 4 通りあることになる．時間 t で，エージェント S と A がとった行動がそれぞれ N, W であるとき，次の時間の状態 x_{t+1} を x_{nw} のように表すと，$x_{t+1} \in \{x_{nn}, x_{nw}, x_{wn}, x_{ww}\}$ である．よって，環境が変動する繰返しゲームの利得行列は**表 3** のようになる．

実際の複雑システムのエージェントは，その環境についての情報をすべて得られるとは限らないので，各時間で表 3 のような利得表が得られることは難しく，したがって，ゲーム理論による解析のように，互いの均衡点を求めることによって自分の行動を決定するのは現実的ではないと考える．むしろ，1 章で述べた背景から，エージェントの行動によって環境が変化する状況では，動的な環境に適応できる能力を持たせる方が現実的であると考える．特に自然の複雑システムの構成員は，各々独自の目標を持ち，自己や環境に評価されながら，試行錯誤で適切な行動を獲得したと考えられる．強化学習は，そのような方式の性格によくあてはまる学習方式である．よって，以後，各エージェントは，このような動的な環境の下で働くことを前提とするが，ゲーム理論による解析ではなく，強化学習によるエージェントが，それぞれの状況で適切な行動を学習する形式で，どのように協調していけるのかを解析する．

3. 実現方法

3.1 シミュレーションの枠組

各エージェントに共通の興味対象の時間 t における状態を x_t とする．x_t は，その対象がとりうる状態集合 X の要素であるから，$x_t \in X$ となる．そのとき，複数のエージェントは，各々独自の目標の状態になるように x を操作したり，影響を及ぼすような行動をとったりしている．

これら複数のエージェントをエージェント A，エージェント B，エージェント C…とする．また，行動の集合をそれぞれ A, B, C, …とし，各エージェントが実際にとった行動をそれぞれ a, b, c, …とすると，$a \in A$, $b \in B$, $c \in C$, …である．

次の時間の状態 x_{t+1} は，現在の状態とこれらの行動によって決まるため，各エージェントの行動集合と，興味対象の状態集合の直積から，興味対象の状態集合への写像となる．その写像を T とすると，

$$T : X \times A \times B \times C \times \ldots \to X \quad (1)$$

となる．本研究では，2 つのエージェントの場合のみを考慮する．また，簡単のため興味対象の状態は 1 次元の数値であり，エージェントが行動した結果，その興味対象の状態に及ぼす影響も数値であるものとする．2 つのエージェントを A，B とし，ある時間にとった各エージェントの行動が及ぼす影響をそれぞれ，e_A, e_B とすると，状態の遷移は

$$x_{t+1} = x_t + e_A + e_B \quad (2)$$

となる．

e_A, e_B は，実数であり，各エージェントの行動 a, b に対して 1 対 1 に対応されているエージェントのパフォーマンスである．

このような状況下では，エージェントは独自の目標に近づけるために，各状況に適した行動を選択する．

3.2 エージェントの実装

つねに変化する環境で活動するには，そのときの状況で最良の行動をとるための学習を行う必要がある．マルチエージェントの問題領域では，強化学習が学習として最も適すると考えられている[7]．強化学習は，学習者が置かれている環境からの感覚入力による状態から行動を決定し，環境からその行動に対して罰や報酬が与えられる．学習者は，報酬が得られるように環境に適応するが，報酬は行動をとった後，すぐには与えられるとは限らず，一連の行動に対して与えられる．

強化学習は，教師あり学習のようにエージェントの入力と出力をつなげる明確な表現はないため，得られる評価によって過去にとった行動の評価だけを得る．すなわち環境についてのモデルが不要である．またオンラインで学習が進行するため，学習が試行錯誤で行われ，動的な環境に適用できるという特徴がある[7)~10)]．

強化学習を実現するには，さまざまな方法が考案されているが，現在の状態とそのときとった行動の組についての評価を行う Q 学習を採用する．具体的には，Q 学習の事例に基づく強化学習法[11)] を採用する．

事例に基づく強化学習を行うエージェントの，各時間の行動の基本サイクルを以下に示す．

(1) 環境からの情報を取得．
(2) どのような行動をとるかを決定する．
(3) 行動する．
(4) 強化信号（報酬，罰）を受け取る．
(5) 記憶の更新

(1) の環境からの情報は，興味対象の現在の状態で

```
var InMem:array[m]ofreal
    OutMem:array[m]ofsymbol
    RifMem:array[m]ofreal
procedure ModifyMem (InData,
                OutData, Rinfc) ;
if CTime≤m then t:=Ctime;
else begin
  forget (InMem,OutMem,RifMem)
  t:=m end
InMem[t]:=InData;
OutMem[t]:=OutData;
if Rinfc=0 then RifMem[t]:=0
else begin
  R:=Rinfc; I:=t;
  while I> 0, |R| > ν and
                RinfMem[I]=0
  do begin RifMem[1]:=R;
  R:=γR; I:=I-1 end
end
```

図 1 事例に基づく強化学習のメモリ更新アルゴリズム
Fig. 1 Memory modifing argorithm for inscetance-based reinforcement learning.

ある．(2) の行動を決定するアルゴリズムは後述する．(4) の強化信号は，それぞれが持つ利得関数によってなされる．つまり，エージェントの行動の結果，興味対象の状態が新しく遷移するが，その興味対象の状態が，各々のエージェントにとってどの程度有利になったか，目標を達成できたかをそれぞれの利得関数によって評価する．

このような意味から，利得関数は環境に必ずしも含まれているわけではない．各エージェントの利得関数は，ユーザが設定することもありうる．環境に含まれるのは，エージェントが興味を持っている対象の状態なのである．

(5) の記憶の更新のアルゴリズムを図 1 に示す．

図 1 の m はメモリの容量，つまり過去何ステップ分記憶するかを限定している．現在のステップ数がメモリ容量を超えると，順次古いものから忘れるようにしている．このメモリ更新アルゴリズムは，ある時間における環境の状態とそのときとった行動の組に対する評価の値を管理する．この組のことを事例という．強化信号として入ってくる評価値は，正か負か 0 であるが，強化信号が 0 以外のとき，過去の事例に対して評価値を割引率 γ で割り引きながら伝播するが，過去の事例にすでに 0 以外の数値が割り当てられていればそれ以上伝播させない．また，過去の事例すべての評価値が 0 であっても，割引した結果の評価値が ν 以下になったら伝播を止める．つまり過去への評価値への伝播は，最大でも，$\log_\gamma \nu$ ステップ前まで遡ることになる．よって，γ も ν も 0 から 1 の定数であるが，γ の方が大きい値となる．

(2) のエージェントの行動の決定は，次のようにして行われる．メモリの中の事例は，過去の環境の状態と，そのときとった行動とその評価が組になっているが，その事例の中で，正の評価値が付加されているものの中から，現在の入力値，つまり現在の環境の状態に最も近い，過去の状態を保持している事例を選ぶ．その事例に保持されている行動が，適した行動として選択される．

もしそのような行動が見つからない場合や，初期状態で何も記憶していない場合は，エージェントがとることのできる行動集合の中から**ランダム**に決定される．

本研究では，環境からの情報として，エージェントが興味を持つ対象の状態 x_t の値のみを感知できるものとする．よって，過去の事例から現在の状態と最も近いものを探す場合は，2 つの数値の距離が最短のものを探すことにする．

エージェントは，このように行動を決定し，2 章で定めた式 (2) に従って興味対象の状態を変更していくことになる．そして遷移した状態が，利得関数で評価されて強化信号を得，記憶を更新し，その新しい状態での行動を決定するというサイクルを繰り返す．

4. シミュレーション

ここでは式 (2) に従って，2 つのエージェントが状態量 x_t に影響を及ぼしてそれぞれの目標値に近づけるシミュレーションを行う．

4.1 目標の違いによるシミュレーション

このシミュレーションでは，各エージェントの行動を，negative extreme (NE)，netative moderate (NM)，none (N)，positive moderate (PM)，positive extreme (PE) として両者同じ行動をとれるものとし，行動による効果はそれぞれ，-3，-1，0，1，3 とする．その他，各エージェントのパラメータは目標値と，それによる利得関数以外すべて同じとする．状態量 x_t の遷移は，式 (2) のとおりで，初期状態は $x_0 = 0$．全部で 2048 タイムステップ分の経過をシミュレーションした．各エージェントのパラメータを**表 4** に示す．

このシミュレーションは各エージェントの目標値を変えて行った．このシミュレーションのねらいは，互いの目標値の差によって，状態量 x_t の遷移にどのよう

表 4 正反対の作用を持つ各エージェントのパラメータ．d_A, d_B は各目標値と x_t との距離 $|x_t - g_A|$, $|x_t - g_B|$.
Table 4 Parameters of two agent which has opposite function. $d_A = |x_t - g_A|$, $d_B = |x_t - g_B|$.

parameter	エージェント A	エージェント B
行動集合	NE, NM, N, PM, PE	
行動の効果	それぞれ −3, −1, 0, 1, 3	
目標値	(g_A, g_B) この値を変えて実行	
利得関数	1 $(d_A \leq 1)$ 0 $(1 < d_A \leq 5)$ −1 $(5 < d_A)$	1 $(d_B \leq 1)$ 0 $(1 < d_B \leq 5)$ −1 $(5 < d_B)$
検出情報	x_t	x_t
記憶容量	50	50
割引率 γ	0.88	0.88
割引下限	± 0.05	± 0.05

な違いがあるかを観察することである．両者の目標値 $(g_A, g_B) = (0,0), (1,-1), (2,-2), (3,-3), (4,-4)$ のときをシミュレーションした．

このシミュレーションは，強化学習を行うエージェントが行動を決めるため，正の強化信号が得られない場合や，シミュレーションを開始したばかりの初期状態では，エージェントは有効な事例を記憶に持っていない．そういうときのエージェントは，行動をランダムに実行して，試行錯誤的に有効な事例を得ようとする．つまり 1 つ 1 つのシミュレーションがつねに同じ結果にはならない．そのため，各目標値で，それぞれ 10 回ずつシミュレーションを行った．

各エージェントの利得関数は，各々の目標値と x_t との距離の関数になっている．この距離が 1 以内だと利得として強化信号 1 が得られ，1 以上 5 以下だと強化信号が 0，5 以上だと −1 となる．つまり，この利得関数によって，どの程度目標値に x_t を近づけることができれば，成功したか，失敗したかなどを決定している．

各エージェントの行動による効果も，その絶対値を同じ値にした．行動は，5 種類あって，それぞれの効果に程度の違いを持たせている．これは，正の報酬を得られるほど目標値の近くに x_t の値を近づけて調整できるようにするためのものである．

このシミュレーションの結果を，x_t の状態遷移で分類すると，大まかに，1）周期パターン，2）一方的適応パターン，3）失敗パターンの 3 つに分かれた．それぞれを**図 2** と**図 3**，**図 4**，**図 5** に示す．

周期パターンは，状態量 x_t が両者の目標値にからみつくように，一定周期で振動するものである．これは，両者がうまく目標を交互に達成できた状況である．この状況のときは，両エージェントとも定期的に正の利得（報酬）を得ることができて，その周期の中では負の利得を得ることがないので，エージェントの行動からランダムさがなくなる．よって，一度この周期的状態におちいると，他のパターンに移行することはない．

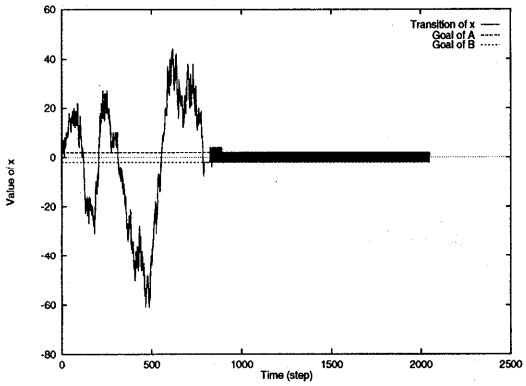


図 2 周期パターン（互恵パターン）．両者がともに目標を達成することができている状態．両者の目標値 $(g_a, g_b) = (2, -2)$
Fig. 2 Cyclic pattern (reciprocal pattern). The situation that both agents can achieve their goal by turns. $(g_a, g_b) = (2, -2)$.

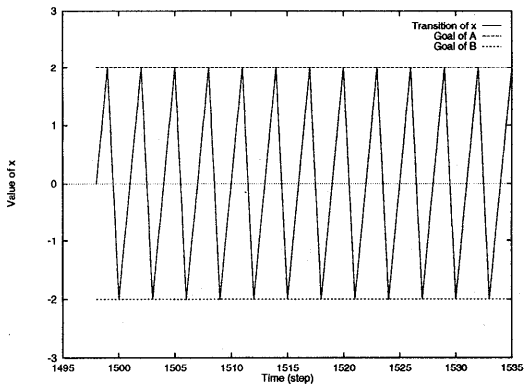


図 3 図 2 の 1500〜1535 ステップの部分．周期的に x が遷移し，安定している状態．$(g_a, g_b) = (2, -2)$
Fig. 3 The behavior of x from 1500 to 1535 step on Fig. 2. The situation that the transition of x is stably periodic. $(g_a, g_b) = (2, -2)$.

一方的適応パターンは，どちらか一方の目標値の近くで x_t が遷移し，もう一方の目標値に近づくことがない．つまりどちらか一方が目標値を達成するよう適応した場合である．このとき，目標を達成しているエージェントのみが正の利得を得ることができるが，もう一方のエージェントは有効な事例を得ることができない状態なので，ランダムに行動する．

失敗パターンは，どちらの目標値も達成することができずに，目標値から遠くはなれたところで x_t が遷移している状態である．このとき両方のエージェントとも 0 か負の利得しか得ることができない．

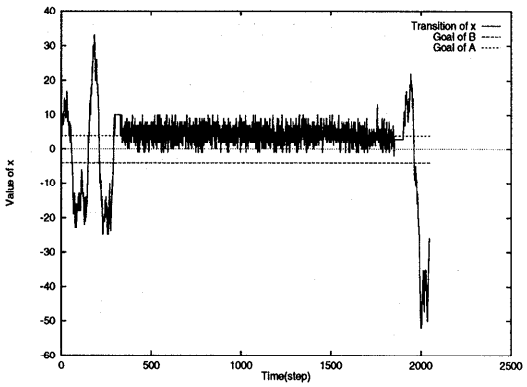


図 4 一方的適応パターン．どちらか一方のエージェントのみが目標を達成できて，もう一方は失敗している状態．両者の目標値 $(g_a, g_b) = (4, -4)$

Fig. 4 One-sidedly adapted pattern. The situation that only one agent can achieve its goal and another fails. $(g_a, g_b) = (4, -4)$.

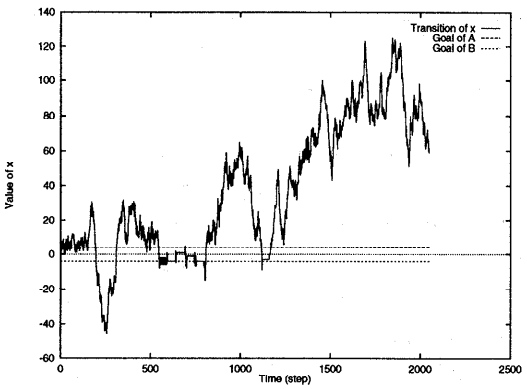


図 5 失敗パターン．両者ともに目標を達成できていない状態．$(g_a, g_b) = (4, -4)$

Fig. 5 Failure pattern. The situation that no agent can achieve its goal. $(g_a, g_b) = (4, -4)$.

各目標値の 10 回のシミュレーションで，それぞれのパターンがどれくらい出現したかを**表 5** に示す．

目標値が $(g_A, g_B) = (0, 0)$ の場合は，両者の利害が完全に一致している場合であるから，自分の利益が相手の利益となる状況である．表 5 から分かるとおり，目標値が一致しているときは，周期パターンになって両者とも利益を分け合うパターンの方が多くなっている．また，目標が一致しているときは，結果のパターンも失敗パターンか周期パターンのいずれかである．

シミュレーションの初期の段階で各エージェントはランダムに行動するが，その間目標値の近くに x_t を近づけられないと報酬が得られない．そのため x_t はランダムに遷移しつづけるが，もしシミュレーション中に目標を達成できなくて報酬を受けとれないと，適応することができずに失敗パターンとなる．

表 5 表 4 の条件で行った利得変化型繰返しゲームのシミュレーション結果．Rp：周期パターン，Op：一方的適応パターン，Fp：失敗パターン

Table 5 The results of the simulations executed under the conditions presented by Table 4. Rp: cyclic pattern, Op: one-sidedly adapted pattern, Fp: failure pattern.

目標値 (g_A, g_B)	Rp	Op	Fp
$(0, 0)$	6	0	4
$(1, -1)$	7	0	3
$(2, -2)$	4	0	6
$(3, -3)$	1	2	7
$(4, -4)$	0	4	6

両者の目標値が $(g_A, g_B) = (1, -1), (2, -2), \cdots$ と，次第に離れている場合は，両者の利害が完全には一致していない状況である．このような状況では，両者の利害が一致している時と比べて，周期的互恵パターンが少なくなっている．また，両者ともに失敗する例も増えている．

両者の目標値が一致しない場合では，どちらか一方のみが目標を達成できる一方的適応パターンが出現している．このパターンは，たとえば，エージェント A だけが一方的に適応できたような場合である．そのとき，エージェント A は，x_t が自分の目標値から負の方向へと離れようとすると positive extreme のような行動をとって引き戻したり，逆に自分の目標値からさらに大きな正の方向へと離れようとすると，negative extreme の行動をとって自分の目標値の近くへと戻すことができるような事例を学習した場合である．このときエージェント B は，適応していないのでランダムに行動するばかりである．

つまり，そのときの x_t の値と適切な行動のペアを，どちらか一方のエージェントが学習することができた状態が，一方的適応状態である．しかしこの状態も必ずしも安定な状態ではなく．適応したエージェントが有効な事例を忘却したり，偶然に相手が x_t を相手側の方へ引き寄せる強い行動をとり続けたりすると，優位性が反転したり，失敗パターンになることがある．

失敗パターンは，図 5 のように，x_t が大きく乱れて遷移するような状況である．失敗パターンの状態からでも，各エージェントの目標値に x_t が近づけば，一方的適応パターンになったり，周期パターンに変化したりすることもある．表 5 で示されている失敗パターンや，一方的適応パターンは，シミュレーション時間 2048 ステップ中大半を占めたパターンである．失敗パターンや，一方的適応パターンから他のパターンへ移行することもあるが，これはすべて適応していないエージェントのランダムな行動によるものである．

表 6 10000 ステップ経過後の状態遷移パターン．Rp：周期パターン，Op：一方的適応パターン，Fp：失敗パターン
Table 6 The results of the behaviors of x_t after 10000 time steps. Rp: cyclic pattern, Op: one-sidedly adapted pattern, Fp: failure pattern.

目標値 (g_A, g_B)	Rp	Op	Fp
$(0,0)$	8	0	2
$(1,-1)$	8	0	2
$(2,-2)$	3	0	7
$(3,-3)$	0	4	6
$(4,-4)$	0	4	6

目標値が離れているときでも，周期パターンは出現している．しかし両者の目標値が $(g_A, g_B) = (4,-4)$ のときは周期パターンは出現していない．

目標値が離れているときの周期パターンは，目標値が一致しているときとは異なっている．各エージェントは，x_t を自分の都合の良いようにもっていこうとしているが，相手もまた同様に行動している．そのようなとき，両者の目標が完全に一致していなくても，相手が報酬を得られる状況でも自分は，負の報酬を受け取らずに，損も得もしない場合に，周期パターンが出現する傾向がある．

エージェントは，x_t の値に応じて正，0，負の報酬（強化信号）を受け取るが，相手が得点をあげているとき，自分は少なくとも損をしなければ互恵パターンになる可能性がある．エージェント A が，0 か正の強化信号を受け取ることができる x_t の領域を V_A，同様にエージェント B のそれを V_B とする．そしてエージェント A，B が得点をあげることができる x_t の領域をそれぞれ G_A，G_B とすると，以下の条件が周期パターンが多く出現している状況の特徴である．

$$V_A \cap G_B \neq \phi \quad \text{かつ} \quad V_B \cap G_A \neq \phi. \qquad (3)$$

ϕ は空集合である．目標値 $(g_A, g_B) = (1,-1)$ のときは，$V_A \cap G_B = [-2,0]$，$V_B \cap G_A = [0,2]$，目標値 $(2,-2)$ のときは，$V_A \cap G_B = [-3,-1]$，$V_B \cap G_A = [1,3]$，目標値 $(3,-3)$ のときは，$V_A \cap G_B = -2$，$V_B \cap G_A = 2$ であるので，これらの目標値のときは，周期パターンが出現しているが，目標値が $(4,-4)$ のときは，これらの共通集合が空なため周期パターンが出現していない．

このことを確かめるため，まったく同様のシミュレーションを，10000 時間ステップ経過させてから 2048 ステップ分の x_t の遷移のパターンを確かめた．結果を**表 6** に示す．

表 6 の結果は，表 5 の傾向と変わらないことが分かる．つまり，目標値が一致しているときは，失敗パターンか周期パターンしか見られず，目標値が離れているときは，互いに妥協できる領域がある場合のみ周期パターンが出現している．

表 7 各エージェントの性質を反映する利得関数．d は各目標値と x_t との距離 $|x_t - g|$
Table 7 Payoff functions reflect each nature of agents. $d = |x_t - g|$.

性質	利得関数	
寛容	1	$(d \leq 1)$
（*tolerant*）	0	$(d > 1)$
普通	1	$(d \leq 1)$
（*normal*）	0	$(1 < d \leq 5)$
	-1	$(d > 5)$
心が狭い	1	$(d \leq 1)$
（*narrow-minded*）	-1	$(d > 1)$

このように，各エージェントは，ランダムな試行錯誤によって目標を達成しようと適応を試みる．その過程で 3 つのパターンが見られるが，両者の目標値の乖離の度合いによって，出現するパターンと出現しないパターンがあることが確かめられた．

4.2 エージェントの性質による違い

エージェントの利得関数は，そのエージェントの目標だけではなく，そのエージェントがどれくらい目標を達成すると，どの程度の成功，あるいは失敗と見なすかを反映していることから，非常に重要である．前章で，エージェントに相手の成功を許す寛容さがあると，周期パターンとなって互いに利益を分け合う可能性が大きくなることに言及した．そこで，ここでは異なった利得関数を実装することで，エージェントの性質に違いを持たせて，各々の目標値でシミュレーションしてみることにする．

そこで，3 種類の利得関数を考える．1 つはこれまでと同じもので，目標値に近ければ正の利得を受け取り，中途半端だと 0，目標値から遠いと負の利得を受け取る．この性質を持ったエージェントを「普通の」（normal）エージェントと呼ぶことにする．2 つめは，目標値に近いと正の利得を受け取るが，それ以外はすべて 0 の利得を受け取る．この性質を「寛容である」（tolerant）と呼ぶことにする．最後は，自分が正の利得を受け取る場合以外は，すべて負の利得値になる利得関数で，この性質を「心が狭い」（narrow-minded）と呼ぶことにする．それぞれの性質とその利得関数を**表 7** に示す．

つまり，これらの性質は，相手の成功に対して寛容かどうかについて分類したものである．心が狭いエージェントは，自分の成功以外はすべて失敗であるし，寛容なエージェントは自分の成功でなくても，別に気にしない．それぞれの 3 種類の性質を組み合わせて

表 8 各エージェントの性質の組合せによるシミュレーション結果．各性質は，N（普通），T（寛容），S（心が狭い）で，結果は，それぞれ，周期パターン/一方的適応パターン/失敗パターン，の数

Table 8 The results of the simulation that the case of each agent's nature is different. Natures are N (normal), T (tolerant), S (narrow-minded). P/O/F shows how many patterns can be seen in it of periodic, one-sidedly adapted, fail pattern respectively.

組合せ	(g_A, g_B)		
(A×B)	(0,0)	(2,−2)	(4,−4)
$T \times T$	26/0/4	18/3/9	12/4/14
$T \times N$	26/0/4	18/2/10	0/5/25
$T \times S$	20/0/10	3/1/26	0/2/28
$N \times N$	21/0/9	12/1/17	0/9/21
$N \times S$	14/0/16	2/0/28	0/5/25
$S \times S$	4/0/26	1/0/29	0/3/27

シミュレーションを行った．組み合わせ方は 6 通りで，それぞれの組合せに，各目標値 $(g_A, g_B) = (0,0)$, $(2,-2)$, $(4,-4)$ の場合をそれぞれ 30 回行った．結果を**表 8** に示す．

結果を見ても，全般的に目標値が離れるにつれて互いが利益を分け合うパターンである，周期パターンの出現回数が減少している．しかしながら目標値が一致している場合でも，エージェントが両者とも心が狭い場合のときには，極端に周期パターンの出現回数が減少している．また，目標値が離れている，$(g_A, g_B) = (4,-4)$ のときでも，両者の性質が寛容であるときは，周期パターンが極端に多く出現している．これらの傾向は，それらの中間である，目標値 $(g_A, g_B) = (2,-2)$ の場合にはっきり現れている．

やはり，互恵的パターンに落ち着くためには，相手の成功に対する寛容さが重要であるという結果になった．また，各シミュレーションで一方的適応パターンも出現したが，どの性質がどの性質に対して優位を占めやすいというような傾向は確認できなかった．

5. 議　　論

本研究では，エージェントに共通の興味対象を導入し，その対象の状態値を各エージェントに都合の良い値へと近づけようと影響を及ぼす場合を考えた．この興味対象の状態遷移を観察することによって，両者ともに利益を得ている状況であるか，一方のみがうまくやっている状況か，両方ともうまくいっていない状況かを確認した．

3 つの状況のどれになるかの傾向をつかむため，それぞれの目標値や性質を変えてシミュレーションを行った．利害が完全に一致している場合を除いて，エージェントの置かれる状況は，両者の目標値と性質の相違によって次のように分類できる．

- 相手が得をするときはこちらが損をするという排他的な状況．
- 相手が得をするときは，自分は損はしないという状況．

本研究のシミュレーションで出現した，周期的パターンは，エージェントがうまく協調している状況であると考える．

文献 5) では，繰返し囚人のジレンマで協調関係を育てる教訓として，「相手を羨望しないこと」，や「自分から裏切らないこと」，「策に溺れないこと」，「相手にされたことをそのままお返しすること」をあげている．本研究のエージェントは，相手についてのモデルを考慮せず，興味対象の状態のみを感知させていたので，最初の 2 つの教訓は，前提に設定されていたといえる．しかし残りの 2 つは，周期パターンになったときの両エージェントの振舞いとまったく同じである．一度 2 つのエージェントが周期パターンとなり，利益を互いに享受し合うように適応すると，それ以上行動に変化は見られなくなる．また，周期的に利益が交互にもたらされるよう振動しているのであるから，これもそのままお返しするという意味にはあてはまっている．

よって，本研究のシミュレーションの周期パターンは，両者がうまくやっていける安定な状況と考えるが，エージェントが動的な環境でうまくやっていくための条件として，4.1 節で示した式 (3) が考えられる．しかし，4.2 節でのシミュレーションでは，式 (3) が成立しなくても周期パターンになる例も見られたため，式 (3) が周期パターンが出現するための条件とはいえないが，式 (3) で示した領域が広ければ広いほど周期パターンになりやすいという傾向には変わりがない．したがって，エージェントが動的な環境でうまくやっていくための条件として次の 2 つを提示する．

(1) 各々の目標値を近づけること．
(2) 他のエージェントの利益になる局面でも，自分は気にせず得も損もしないような余裕を持たせること．

自然界には，全体として恒常性を維持する複雑なシステムが数多くある．たとえば，生体の免疫システムもそうである．これは，免疫機能を賦活させる要素と抑制する要素からなっており，それらの要素も免疫機能の強さからのフィードバックを得て制御されているというネットワーク構造になっている．つまり両者は，免疫機能の強さを同時に操作することで，互いに影響を及ぼし合っているわけである．

両者が互いに他者を認める余裕を持っていれば，免疫機能が正常に働いていると考えられるが，たとえば，アレルギー反応のように，免疫システムが異常をきたして，賦活要素に寛容さがなくなり，自分以外のものすべてを排除しようとする過剰な反応を示す場合もある．このような場合は，心の狭いエージェントによって状態遷移が不安定になってしまう場合に相当すると考えられる．

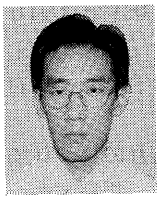
1章で，恒常性を維持するシステムでは，各エージェントの行動は，他の構成員から容認されるようなものであることを予測した．今回のシミュレーションによって，各構成員の行動が環境を変動させ，それが他者にも影響を及ぼすような状況では，構成員同士の目標値が一致しているか，他者に対する寛容さがある場合には，環境の状態も安定に遷移することが確認できる結果が得られた．逆に，目標値が離れていたり，他者に対する寛容の度合いがない場合は，環境の遷移は不安定になることも同時に確かめられた．

本研究では，利得値が変化するゲームで動的な環境を設定し，そこで必然的に要求される適応能力を，反射型の強化学習能力として2つのエージェントに実装した．しかし，エージェントのアーキテクチャを異なったもの，たとえば，熟考型と反射型の強化学習能力を持ったエージェントがそれぞれ対峙するシミュレーションも可能である．さらに，ここでの学習結果をどのようにして用いるのかということも，今後の課題である．

参考文献

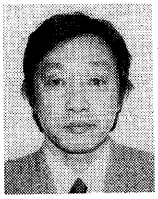
1) 有馬　淳：複雑系としてのエージェント社会，人工知能学会誌，Vol.13, No.1, pp.5–6 (1998).
2) 松原繁夫，横尾　真：繰り返しゲームにおいて協調行動を生成する先読み型行動選択方法，人工知能学会誌，Vol.12, No.6, pp.881–890 (1997).
3) 沼岡千里，大沢英一，長尾　確：マルチエージェントシステム，共立出版，pp.19–47 (1998).
4) 鈴木光男：新ゲーム理論，勁草書房 (1994).
5) Axerlrod, R., 松田裕之（訳）：つきあい方の科学，HBJ出版局 (1987).
6) Pollack, M.E., Joslin, D., Nunes, A., Ur, S. and Ephrati, E.: Experimental Investigation of an Agent Commitment Strategy, Technical Report, Department of Computer Science and Intelligent System Program, University of Pittsburg, Pittsburg (1994).
7) 三上貞芳：強化学習のマルチエージェント系への応用，人工知能学会誌，Vol.12, No.6, pp.845–849 (1997).
8) Kaelbling, L.P. and Littman, M.L.: Reinforcement Learning: A Survey, *Artificial Intelligence Research*, Vol.4, No.5, pp.237–285 (1996).
9) 山倉雅幸，宮崎和光，小林重信：エージェントの学習，人工知能学会誌，Vol.10, No.5, pp.683–689 (1995).
10) 畝見達夫：強化学習，人工知能学会誌，Vol.9, No.6, pp.830–836 (1994).
11) 畝見達夫：事例に基づく強化学習，人工知能学会誌，Vol.7, No.4, pp.697–707 (1992).

(平成10年 7月10日受付)
(平成10年12月 7日採録)



稲垣　裕伸（学生会員）

1971年生．1996年室蘭工業大学大学院工学研究科博士前期課程情報工学専攻修了．現在同大学大学院工学研究科博士後期課程生産情報システム工学専攻．人工知能学会学生会員．



魚住　　超（正会員）

1973年室蘭工業大学電子工学科卒業．その後，北海道大学において1980年に工学博士を取得．カナダ・サスカチワン州立大学研究員，国立特殊教育研究所を経て，1988年から室蘭工業大学情報工学科の助教授となり現在に至る．研究分野は，時系列信号，画像，機械学習医用福祉工学領域への応用．人工知能学会，IEEE，電子情報通信学会，計測自動制御学会，日本ME学会各会員．

小野　功一

1938年生．1964年北海道大学医学部医学科を卒業．1969年同大学大学院医学研究科卒業．同大学医学部講師を経て，現在室蘭工業大学情報工学科教授．医学博士．専門研究分野は呼吸力学，循環力学．