社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－ICS－141 (8)
2005／11／9

非零和ゲームの強化学習—相手の行動を読むプログラム

伊藤 昭† 大橋資紀† 寺田和憲†

利害の対立する状況で，相手の心を読んで適切に行動するプログラムを目標に，「繰り返し非零和ゲーム」を計算機に解かせることを試みる．具体的には，高得点を得るためには相手の心を読むことが必要な問題として「1・2・5じゃんけん」というじゃんけんの一種を取り上げ，この問題を人はどのように解くのか，どのようにすれば強いプログラムが作れるのかを検討した．次に，過去 h 個の履歴を用いて次の手を決定する k 次マルコフ戦略という概念を導入し，履歴を用いたQ学習を定式化した．また，履歴を用いたQ学習に1・2・5じゃんけんを学習させることにより，プログラムが協調解を学習できることを示し，協調までの時間の振る舞いを調べた．最後に，このゲームに対する人や学習プログラムの振る舞いから，心を読むと言えるより強いプログラムの作り方を提案した．

Reinforcement Learning of Non-Zero-Sum Games A Program That Reads Other's Actions

Akira Ito,† Motoki Ohashi† and Kazunori Terada†

We make a computer solve "iterated non-zero-sum games", with the aim to develop an algorithm to read other's mind and act appropriately under the competitive situation. We take up "1-2-5 Janken", a variant of "paper-scissor-rock" game, and investigated how humans solve it, and how to make a strong program for computers. For that purpose, we introduced k-th Markov strategy, and formalized Q-learning based on history. Applying the above to 1-2-5 Janken, it is shown that a program can learn "cooperative strategy." Lastly, from the examination of the behavior of humans and programs, we propose a method to develop a program which reads others' minds (behaviors).

1. はじめに

人は相互に利害の対立する状況下では，お互いに相手の心を読み，少しでも自己に有利な結果を得ようと努力する．しかしながら，これまで機械は人に奉仕するものと考えられ，人と機械との利害が対立するということはあまり想定されてこなかった．しかしながら，「鉄腕アトム」でも取り上げられているように，人の中に利害の対立がある以上，その反映として人と機械の間に利害の対立が生じることは避けられない．たとえば，機械（ロボット）が利害の対立する人のエージェントとして行動しているとき，人はその機械が常に自分のために行動してくれると期待するのは危険である．

たとえ機械同士であっても，それらが利害の対立する人のエージェントである場合など，利害の対立下でインタラクションしなければならない状況は当然生じ得るであろう．さらに言えば，嘘や駆け引きの無い（禁止されている）社会と，人間社会のようにそれらの可能な社会とでは，どちらが協調的タスクをうまくやり遂げられるのかと考えたとき，常に前者の方が有利とは言えないのではなかろうか．「嘘も方便」と言う言葉があるように，嘘や駆け引きがあることで利害の調整がうまくいく可能性もあるわけである．

このように考えると，機械（ロボット）にも相手の心（行動）を読んで「駆け引き」をする能力があれば，人に代って進出できる領域は大きく拡がるものと思われる．また，実際に機械が積極的に駆け引きを行わなくても，人の「駆け引き」の行動が理解できるようになれば，人と機械のインターフェースもまた違ったものになるのではないかと思われる．

このような期待を持って，我々は機械（計算機）に相手の心（行動）を読む能力を付与するための研究を行っている．ここでは手始めとして，高得点を獲得するためには相手の心を読むことが必要となると思われる「繰り返し非零和ゲーム」を，計算機に解かせることを試みる．

繰り返し非零和ゲームは，我々の社会のインタラク

† 岐阜大学工学部
Faculty of Engineering, Gifu University

ションのモデルとして，古くから様々な分野の研究者により研究がなされてきた．なかでも囚人のジレンマゲームは，Axelrod らが公開のゲームコンテストを行ったことで，広い分野の研究者の関心を集めることとなった．囚人のジレンマゲームについては多くの文献があり[1)2)3)]，ここでは詳しくは述べない．しかしながら我々が残念に思うことは，しっぺ返し戦略 (Tit for Tat) が他の様々な戦略を抑えてコンテストで優勝したことにより，このようなゲームにおいて「相手の心を読む」ことがあまり有効ではないという印象を多くの人に与えてしまったことである．これに対して我々は，囚人のジレンマゲームを解くにあたって，相手の心を読むことが重要であることをシミュレーションを用いて示している[4)]．

我々の日頃の行動を反省してみると，人との付き合いにおいて単純な「しっぺ返し」だけでうまくいかないことを我々は良く知っている．実際に我々は，相手に応じて強く出たり，妥協したり，またある方向に誘導したりと，目的実現のために様々な戦略を駆使している．しかしながら，純粋にアルゴリズムとして考えたとき，我々が人との付き合いにおいて用いている「相手を読む」という戦略は，利害の対立する場で本当に有効なのだろうか．もしそうだとするなら，その論理・アルゴリズムはどうなっているのだろうか．そのようなアルゴリズムを計算機に学習させることはできないものだろうのか．本論文は，そのような方向での研究の第一歩である．

2. 繰り返し型対称非零和２人ゲーム

もし人が，非零和ゲームを同じ相手と繰り返し対戦せねばならないとき，得点の期待値を最大にするためにはどのような行動を取ったら良いのだろうか．たとえあらゆる状況で「最適」という戦略は存在しなくても，相対的に良い (強い) 戦略というものは存在し，その共通の性質が議論できるのではないだろうか．実は，様々な未知の相手と対戦して良い性能を発揮するためには，相手の心（行動）を読むことが必要となるはずだ，というのが我々の期待である．

まず，ここで扱う対称非零和ゲームについて説明する．零和ゲームとは，自分と相手の利得の和が常に零となるゲームで，勝敗のみが問題となる対戦ゲームはこの範疇にはいる．零和ゲームでは，相手の利益と自己の利益が完全に相反するので，相手の最適行動が自己にとって最悪の行動となり，相手の心を深く読むことなく相手の行動の予測が行える．一方非零和ゲームでは，相手の利得に対する情報が簡単には得られないため，行動の予測はより困難となる．

対称ゲームとは，相手の利得行列と自己の利得行列とが同じ構造をしているもの，すなわち相手と自分との立場を入れ替えてもゲームの構造が変わらないものである．ここでは問題を対称ゲームに限定することで，相手の利得行列に対する情報を利用可能とするが，非対称ゲームの場合でも何らかの方法で相手の利得情報を獲得できれば，ほぼ同じ議論が展開できる．

繰り返し対称２人非零和ゲームを以下のように定義する[5)]．まず，次のような対称２人非零和ゲームが与えられている．各プレーヤの取り得る戦略の集合を $\Sigma = \{s_1, s_2, \ldots, s_n\}$ とする．プレーヤ P1, P2 の戦略を $s^1 = s_i$, $s^2 = s_j$, P1 の利得を $f^1(s^1 = s_i, s^2 = s_j) = g_{ij}$ とすると，P2 の利得は，対称性から $f^2(s^1 = s_i, s^2 = s_j) = f^1(s^1 = s_j, s^2 = s_i) = g_{ji}$ となる．g_{ji} は行列として表現できるため，利得行列と呼ばれる．

ここでの目的は，このゲームを N 回繰り返して行い，その間の総得点を最大化する政策（行動戦略）π を設計することである．ただし，N は長期的な作戦を立てることが有効な程度には十分大きい (たとえば $N \geq 10^6$) ものとする．政策は N 回の繰り返しゲームを一つのゲームとして考えれば「戦略」であるが，一回毎のゲーム（S ゲームと呼ぶ）の戦略 s と区別するために，繰り返しゲームの戦略 π を「政策」，また S ゲームの戦略 s を「手」と呼ぶことにする．

政策 π は，これまでに出された相手の手，自分の手の履歴を用いて，次に自分の出すべき手 $s^m(t)$ を決定するアルゴリズムである．ここで，$\{s(k)\}_i^n = \{s(l), s(l+1), \ldots, s(n)\}$ という表記法を導入する．これにより，政策 π は時刻 k における自己，および相手の手を $s^m(k), s^o(k)$ として，

$$s^m(t) = \pi^m(\{s^m(k)\}_1^{t-1}, \{s^o(k)\}_1^{t-1}, t)$$

となる関数 π^m として定義される．

ゲーム理論では，双方とも自己の手が相手の手に対する最適戦略（手）となっているとき，この手の組を Nash 均衡解と呼ぶ．一般の利得行列では，このような Nash 均衡解 (s^m, s^o); $s^m, s^o \in \Sigma$ が必ずしも存在するわけではない．そこで，次に出すべき特定の手 $s \in \Sigma$ を指定するのではなく，次に出すべき手を適当な確率 $\pi(s_i)$ で $s_i \in \Sigma$ の中から選択するように戦略の定義を拡張する（混合戦略と呼ばれる）と，その中では Nash 均衡解が必ず存在することが知られている．

戦略が確率でしか与えられないのは，評価関数の不確実さに起因するのではなく，相手からの「読み」を拒否するためであり，他の知的エージェントが存在す

るゲーム理論的状況での最適行動の生成には必須の機能である．そこで我々は，強化学習の枠組みに以下のような確率政策を導入する．確率政策 π とは，これまでに出された相手の手，自己の手の履歴を用いて，次に自分の出すべき手の確率を決めるアルゴリズムである．

混合戦略も含めて双方の戦略の組を決めたとき，その利得 (の期待値) が，一方の利得を下げずに他方の利得を上げ得ないとき，言い換えれば両者を同時に満足させる戦略の変更が不可能なとき，この戦略の組合わせをパレート (Pareto) 最適と呼ぶ．零和ゲームでは全ての戦略の組が定義からパレート最適である．しかし，非零和ゲームでは，状況はそれほど簡単ではない．繰り返しゲーム，S ゲームいずれについても，利得表によってはパレート最適な Nash 均衡解が存在しないことがある．

S ゲームではパレート最適な Nash 均衡解が存在しない場合でも，繰り返しゲームでは状況が変わる可能性がある．というのも，対戦を繰り返すことによって，個々の対戦では存在しなかった Nash 均衡解が生じる可能性があるからである．実際，S ゲームでは Nash 均衡解がパレート最適ではないときにこそ繰り返しゲームが問題になる．というのも，S ゲームにおける戦略を適当に組み合わせることで，双方が納得できるパレート最適な解を見いだすことができれば，共に利益を得ることができるからである．

3. 1・2・5じゃんけん

「1・2・5じゃんけん」と我々が呼ぶものは，じゃんけんの一種であるが，勝敗だけが争われるのではなく，グー，チョキ，パーそれぞれで勝つと 1, 2, 5 点が勝った人の得点になるというものである．子供のころ神社の石段などで，「じゃんけんをしてグー，チョキ，パーで勝つとそれぞれ勝った方が 1, 2, 5 段上がる」という遊び (または類似のもの) をやられた記憶はないだろうか．さてこれには，負けた人が得点を払うという零和バージョンと，まけた人は損をしないという非零和バージョンが考えられる．石段登りで言えば，負けても石段を下りなくても良ければ (普通はこのバージョンだと思うが) 非零和バージョンである．もし，負けた方がその分だけ下りなければならないのが零和バージョンである．ここで取り扱うのは，負けても得点を支払わない非零和バージョンである．

1・2・5じゃんけんは，形式的には，グー (G)，チョキ (C)，パー (P) により次の利得表で与えられる対称２人ゲームとして定義される．対称ゲームであるので第一プレーヤ (縦側プレーヤ) の得点のみが記述されている．

	G	C	P
G	0	1	0
C	0	0	2
P	5	0	0

表 1 1・2・5じゃんけんの利得表

ここで，自己が G, C, P を確率 x, y, z で，また相手が G, C, P を確率 g, c, p で出すとすると，自己，および相手の得点の期待値 f^m, f^o は以下で与えられる．

$$f^m = 5zg + xc + 2yp$$
$$f^o = yg + 2zc + 5xp$$

Nash 均衡解は，相互に相手の手の最善手になっている戦略の組として定義され，次のようにして求めることができる．相手が G, C, P を出したときの相手の得点は $5x, y, 2z$ であるから，これを等しくする $(5x = y = 2z)$ 戦略を考えると，

$$(x, y, z)_N = (2/17, 10/17, 5/17)$$

を得る．こちらがこの手を出すことにより，相手がどのような手を出そうと相手は得点 $f^o_N = 10/17$ を得ることになる．ここで，相手も同じ行動を取れば，こちらの得点も $f^m_N = 10/17$ となる．この状態では双方とも自己の戦略を変化させても，もはや得点を改善することはできない．したがって，これがこのゲームの Nash 均衡解となる．

お互いに相手に勝とうとするのではなく，何らかの方法で協調して戦略を調整できるのであれば，お互いに C を出すのをやめるだけで

$$f^m_c = f^o_c = 5/4,$$
$$(x, y, z)_c = (g, c, p)_c = (1/2, 0.1/2)$$

という Nash 均衡解より高い得点を得ることができる．これは常に P を出すという戦略に支配される（得点で優位に立たれる）ため，Nash 均衡解ではない．もっと単純に両プレーヤがこのゲームをランダムにプレイするだけで，平均得点は 8/9 となり，Nash 均衡解の得点 10/17 を上回ることができる．

ここまでは，まだ毎回決められた確率で G,C,P の手を出すという S ゲームでの戦略であるが，もし何ステップかにわたって相互に手を調整できるのであれば，別の戦略も可能となる．たとえば，双方が交互に G, P を出すことで，平均として $f^m_{c2} = f^o_{c2} = 5/2$ を得ることもできる．実は無限回の繰り返しゲームでは，この戦略がパレート最適な Nash 均衡解となる．

このように多様な戦略が可能であり，それぞれの戦略についてどれが有利かは相手の戦略次第である．では，このようなゲームを未知の相手と対戦せねばなら

ないとき, エージェントはどのように行動すれば良いのだろうか. 我々は「最適」という言葉がゲーム理論ですでに様々に定義されていることから, 混乱を避けるためにこのような状況で高い得点を得られる戦略を「強い」戦略と呼び, 強い戦略を構成する方法を検討する. これは Shoham らの best responce と類似した発想である[6].

4. 人は繰り返し非零和ゲームをどのように戦うのか

我々は, 非零和ゲームを戦う「強い」アルゴリズムを求めたいのであるが, その前に人はこのような問題をどのように戦うのかを調べてみることにする.

実験概要

我々は, 繰り返し非零和ゲームでの人の戦い方を調べるため, 次のような実験をおこなった. 具体的には, 1・2・5じゃんけんを 100 回繰り返す実験を 26 組行った. その内の 10 組は, 被験者同士の対戦であり, 他の 16 組は, 被験者と計算機による対戦である. さらに後者では, 16 組の内の 11 組には, 被験者に対して「相手は同じように実験に参加している人である」との教示を, 残りの 5 組に対しては「相手は計算機である」との教示を行った. 実験では, 被験者はそれぞれ別の部屋に入り, ネットワークを介して対戦をしてもらった.

実験を始める前に, 「目的は自分自身の得点をできるだけ高くすることであり, 相手との得点の差を争うことではない」との教示を行った. このことを明確に示すため, 得点に応じて報酬を支払うこととした. 具体的には, ランダムに手を選択した場合の平均得点が 8/9 であるため, 一回のじゃんけんにつき参加費を-1 点とし, 正の得点に対して得点に比例した報酬を支払った. じゃんけんは, 20 回で 1 ゲームとし, 連続して 5 ゲーム 100 回の対戦を行った.

被験者と計算機との対戦では, 計算機の対戦アルゴリズムには次の N 戦略と名付けたものを用いた.

N 戦略

(1) 前回相手に P で勝ったら, 次は G を出す.
(2) 前回相手に G で負けたら, 次は P を出す.
(3) それ以外の状態では G,C,P を Nash 解の割合 (G=2/17:C=10/17:P=5/17) で出す.

これは単純なアルゴリズムではあるが, 協調のパターンが現れたら, それ以後は協調を続けようとするアルゴリズムとなっている.

被験者同士の対戦

被験者同士の対戦結果を表 2 に示す. 協調することを思いついた被験者もいたが, 最終的に協調解に達した組はいなかった. No.1, No.2 の組では一方が協調を試みたが, 相手に搾取されるだけの結果となり, 最終的に協調を諦めている. 協調の試みが失敗した例として No.2 の得点推移を図 1 に, 協調が生じなかった例として No.4 の得点推移を図 2 に示す.

表 2 P1:被験者, P2:被験者対戦による得点

No	P1 得点	P2 得点	No	P1 得点	P2 得点
1	-37	43	6	-49	-5
2	45	-33	7	-19	-33
3	-9	-30	8	-24	-26
4	-41	-34	9	-27	2
5	-31	-17	10	-9	-35

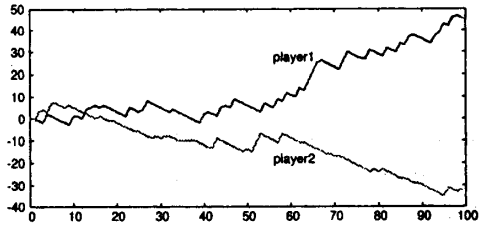

図 1 協調の試みが失敗した場合の得点推移

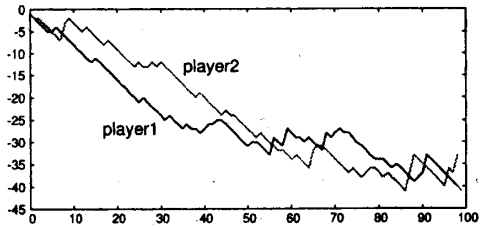

図 2 協調が生じなかった場合の得点推移

被験者と計算機の対戦

被験者と計算機との対戦結果を表 3 (相手は被験者であると教示), 表 4 (相手は計算機であると教示) に示す. 相手を人と教示した場合では, 11 組中 3 組で明確な協調が見られ, 正の得点を獲得している. 残り 8 組の平均得点も-26.9 である. 一方, 相手が計算機だと教示した場合では, 被験者はいずれも協調行動をとることがなく, その平均得点は-40.0 である. それぞれの場合について, 被験者 (P1) の得点の推移を図 3, 図 4 に示す.

人は, 行動の履歴のみを用いて相手モデルを作るわけではない. 事前に相手モデルを構築した上で, それを履歴を用いて修正していく. この事前モデルは, 相手が人である場合と, 計算機である場合では異なったものとなり, その結果相手によって戦略が変化するこ

表 3 P1:被験者,P2:計算機（相手は被験者と教示）対戦による得点

No	P1 得点	P2 得点	No	P1 得点	P2 得点
1	-21	-15	7	-2	0
2	-40	-7	8	-31	-6
3	-28	-3	9	-21	-26
4	56	71	10	56	72
5	-46	-25	11	-26	4
6	16	14			

表 4 P1:被験者,P2:計算機（相手は計算機と教示）対戦による得点

No	P1 得点	P2 得点	No	P1 得点	P2 得点
1	-33	-17	4	-49	-41
2	-36	1	5	-31	6
3	-50	-5			

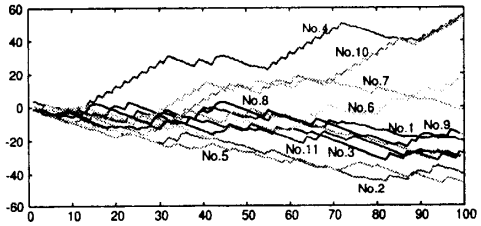

図 3 P1:被験者,P2:計算機（被験者と教示）対戦：P1 の得点推移

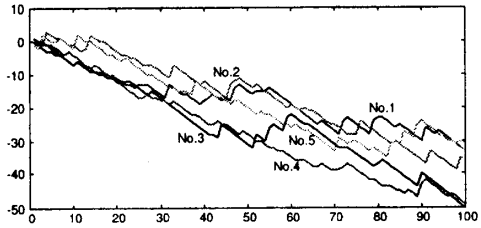

図 4 P1:被験者,P2:計算機（計算機と教示）対戦：P1 の得点推移

とになる.

相手モデルの修正でも, 相手が人と計算機の場合では異なった手順が取られると思われる. 相手が人の場合, 相手の行動パターンを観測し, その行動の意味を意図のレベルで理解した上で, 相手モデルを修正する. したがって, 協調の意図をくみ取り, 相手と協調することが可能となる.

一方相手が計算機の場合は, 人は発見したパターンの意味を意図のレベルで理解しようとはしない. そのため, たとえ計算機が人と同じように協調しようとしても, 人はそこに協調の意図を発見することができず, 相手の手を単なる統計パターンとして見てしまい, その結果自分も協調的行動がとれず, また得点も低くなってしまう.

相手が人の場合でも, 被験者同士の対戦では協調が起こらなかった. これは, お互いが相手の手から協調の意図を汲み取れなかったからだと考えられる. そうなると, 自分の方から特定の (協調を誘う) パターンを出す意味がなくなり, 最終的に, 自分の出す手の割合を最適化するという Nash 戦略に近い戦略に落ち着いたのだと思われる.

被験者の一部に協調しようという動きは生じたものの, 現実には相手に利用されるだけで終わっている. 協調を誘うためには,「こちらは協調の意思がある. しかし協調が成立しなければこちらは最強の手段で反撃する.」というメッセージを送らなければならないのかも知れない. もしそうであれば, たとえ協調ということに気づいた被験者であっても, このようなメッセージをじゃんけんの手を用いて伝達する方法を思いつかなければ, 協調を諦めざるを得ないであろう.

結果を要約すると次のようになる. 人同士の対戦では, 100 回という短い時間では協調行動は生じなかった. しかしながら, 相手が人だと教示され, 且つ相手の戦略が一貫して協調を誘うようなものであれば, 一部の人はその意図に気づいて協調行動をとることができた.

5. 履歴を用いた強化学習

次に, 強化学習をするプログラムがこの問題をどのように解くかを調べてみることにする. そのため, 以下では繰り返し非零和二人ゲームを想定して, 履歴を用いた Q 学習を定式化する.

繰り返し非零和二人ゲームでは, 二人が同時に手 $(a_1, a_2), a_1 \in A_1, a_2 \in A_2$ を出し, その結果利得 $r_1(a_1, a_2)$, $r_2(a_1, a_2)$ を得る. 一般に政策は, これまでの自己および相手の手の履歴を用いて, 次に自己の出す手 (またはその確率) を決める関数 π となる. しかしながら, 任意の政策を考えようとすると, その空間は巨大なため, 探索の対象を次の k 次マルコフ政策（戦略）に制限する.

k 次マルコフ政策 k 次マルコフ政策 π とは, これまでの全ての手に依存するのではなく, 時間普遍な形で過去 k 回の手のみに依存する政策である. すなわち, 時刻 t に自己の出す手が a である確率 $p(a,t)$ が

$$p(a,t) = \pi(a|\{a^m(\tau), a^o(\tau)\}_{t-1}^{t-k}),$$
$$\{a^m(\tau), a^o(\tau)\}_{t-1}^{t-k} = \{a^m(t-1), a^o(t-1),$$
$$a^m(t-2), a^o(t-2), \ldots, a^m(t-k), a^o(t-k)\}$$

で記述される政策である. 相手が k 次マルコフ政策をとっているのであれば, 状態 S_t を

$$S_t = \{a^m(\tau), a^o(\tau)\}_{t-1}^{t-h}$$

と過去 $h(h \geq k)$ 回の履歴で定義することで, 相手エージェントを含んだ環境をマルコフ決定過程（MDP）としてモデル化することができる.

学習政策 学習政策とは, 少数の内部状態 $\{Q_i\}$ を含む k 次マルコフ政策として記述できるものである. ただし, $\{Q_i\}$ が過去全ての手に依存して変化することを認めるので, 学習政策は正確には k 次マルコフ政策ではない. しかしながら, この内部状態は「ゆっくりと変化する」ものと考え, 学習政策を短期的にはあたかも k 次マルコフ政策のように扱う. このとき, 内部状態 $\{Q_i\}$ の変化を「学習」と表現する. 逆に, 変化する内部状態を含まない k 次マルコフ政策を k 次マルコフ固定政策と呼ぶ. 固定政策も相手の過去の履歴を使って自己の手を変え得る確率政策であることに注意をされたい. 以後, 可能な政策を (あらかじめ次数 k を固定しておいた上で) ここで述べた意味での k 次マルコフ学習政策に限るものとする.

相手の政策を k 次マルコフ固定政策, また状態を過去 $h(h \geq k)$ 個の履歴とする. 自己の政策を π として, $Q^\pi(S, a^m)$ を, 状態 S で自分が手 a^m を選んだときに得られる γ で割引された報酬の期待値とする. このとき, 自己の政策 π が最善の (= Q 値を最大化する) 政策であるとすると,

$$Q^\pi(S, a^m) = \sum_{a^o} p(a^o|S)(r(a^m, a^o) + \gamma \max_a Q^\pi(S'(a^m, a^o), a))$$

の関係を得る. ここで $S'(a^m, a^o)$ は, S から行動 (a^m, a^o) によって遷移した状態, $p(a^o|S)$ は状態 S での相手の行動が a^o である確率 (相手の戦略) である. 逐次近似法で Q の値を求めることにすると, 次の式を得る.

$$Q(S, a^m) \leftarrow (1-\alpha)Q(S, a^m) + \alpha(r(a^m, a^o) + \gamma \max_a Q(S', a))$$

行動選択は ε-greedy と Boltzmann 型とを組み合わせて行う. すなわち, 現在の状態を S とすると, 行動選択には以下のアルゴリズムを用いる.

・ε の確率でランダムに $a \in A^m$ を選択
・それ以外

確率 $p(a) = C\exp(Q(S, a)/T)$ で $a \in A^m$ を選択

なお, ここで $\varepsilon(0 < \varepsilon 1)$ はランダムな行動の割合を制御し, $T(T > 0)$ は探索への許容度を決めるパラメタである. 以下ではここで述べた政策を履歴長 h を用いて SQh(履歴長 hQ 学習) と表す.

6. 強化学習プログラムの対戦

前節で述べた履歴長 hQ 学習 SQ$h(h = 0, 1, 2, 3)$ と Random 戦略とをリーグ形式で対戦させた時の得点の平均値を表 5 に示す. 表の得点は左側の戦略が, 上側の戦略と対戦したときの 1step 辺りの平均得点である. 学習に使用したパラメタは, $T = 0.2, \alpha = 0.1, \gamma = 0.9, \varepsilon = 0.01$ である. 実験では 10^8 回の対戦を行い, 表の値はそのうちの最後の 10^7 回の平均値である. さらに, 同じ実験を乱数を変えて 10 回行い, その平均を取っている.

表 5 履歴を用いた Q 学習のリーグ戦での平均得点

	Random	SQ0	SQ1	SQ2	SQ3
Random	0.889	0.673	0.673	0.671	0.670
SQ0	1.631	0.897	0.572	0.554	0.589
SQ1	1.617	1.149	1.958	2.225	2.076
SQ2	1.610	1.123	2.272	2.298	2.371
SQ3	1.599	1.256	2.148	2.333	2.404

ランダム戦略と学習戦略とが対戦すると, 学習戦略は常に P を出すことを学習する. 学習戦略同士では, 履歴長 0 の PQ0 はどの戦略とも協調できず, 履歴長の長い戦略との対戦では一方的に搾取されるだけである. 履歴長 1 以上の学習エージェント同士では, 各エージェントは相互に協調することで高得点を得ることができる.

Q 学習エージェント SQ2 同士の対戦での, 平均得点の時間変化を図 5 に示す. 学習に用いられたパラメタは, 上と同じである. グラフは 1 回の試行のもので, 得点は時間間隔 10^3 で平均している. なおこの実験では, P1 の学習を 5×10^7step で停止 $(\alpha = 0)$ している.

図 5 では, 履歴を用いた Q 学習同士の対戦において, 平均得点は約 2.5 点を上限として激しく振動しているが, これは $\varepsilon > 0$ が原因である. パラメタ ε は Q 値にかかわらずエージェントがランダムに行動する確率を表しており, 一方がこれにより協調からはずれた行動をとると, その時点で協調は崩壊する. しかしながら, 協調が崩れた状態での平均得点は極めて低いため, その後の探索によりまた協調行動が復活するということを繰り返している.

一旦協調が成立すれば $\varepsilon = 0$ とすれば良いという考え方もあるが, 我々はそうは考えない. $\varepsilon = 0$ は探索を放棄することを意味し, その結果それ以後相手が変化したときに適切な対応をする能力を失う. ε の効果は, 図 5 で P1 が学習を停止した場合の P2 の振る舞いに見ることができる. P1 が学習を停止しても, それだけで協調が崩れるわけではない. しかしながら, それ以後に P2 が ε により協調をはずしたとき, P1 はもはや適切な反撃をできない. それを発見した P2 は, 自分により有利な安定解に移行することになる. このようなことが可能なのは, P2 が ε により探索を継続しているからである.

図 6 は, 同じ履歴長を持つエージェント同士 (履歴

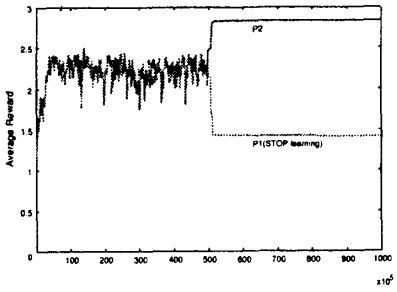

図 5 Q 学習エージェント SQ2 同士の対戦の様子

長 $h = 1, 2, 3$) で, パラメーター ε を変更してを対戦させたものである. ε 以外のパラメーターは, 上と同じものを使用している.

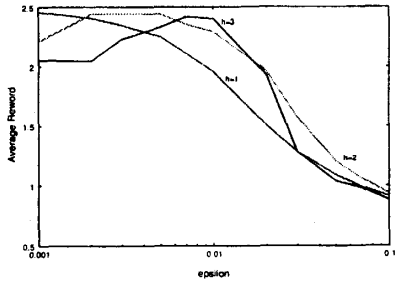

図 6 ε を変化させた場合の平均得点の変化

ε が大きくなると, 探索の速度は早くなるが, 協調を崩壊させる頻度も高くなる. 図 6 から, それぞれの履歴長に適切な ε があることが分かる. 履歴長を長くすると, より大きな ε で平均得点が最大になる. これは履歴長の長さが協調解の安定性に寄与しているからと考えられる.

7. 協調成立までの時間

前節で, 履歴を用いることで Q 学習も協調行動をとれることを確認した. では, この結果から「計算機でも人のように協調できる」と言えるかというと, 問題が残る. というのも, 協調までに要する時間がかかりすぎるのである. 我々は, 被験者を用いた実験で, 協調が生じる場合はそれが数十 step で可能であることを見てきた. それに対して Q 学習同士の対戦では, 協調が生じるのに必要な時間は $10^5 \sim 10^6$ 程度であり, 人の場合と本質的に異なっている.

そこで, このことを詳しくみるために, 前節で述べた履歴を用いた Q 学習を中心に, 幾つかのアルゴリズムを対戦させ, 協調実現までの時間を調べてみた. 使用したアルゴリズムは次のものである.

(1) 履歴長 1 の履歴を用いた Q 学習 SQ1(Q)
(2) 4 節で説明した N 戦略 (N)
(3) Q 学習を協調に誘導することを目指した FCAQ(Fitted for Cooperation Against Q Learning) 戦略 (FCAQ)
(4) SQ1 同士の対戦で協調行動を獲得した学習後の戦略 (LearnedQ)
である.

なお, FCAQ のアルゴリズムは以下の通りである.

FCAQ 戦略 ステップ毎に, 状態, 相手の行動, 相手の得点から Q 学習と同様の計算式を用いて, 相手 Q 値と行動確率を推定する. それを基に次の規則で自身の行動を決定する.

1. 前回 P で勝利のとき G
2. 今回相手が G を出す確率が 0.33 以上のとき P
3. 今回相手が P を出す確率が 0.33 以上かつ次回相手が G を出す確率が 0.33 以上のとき G
4. 上記以外のとき C

図 7 は, それぞれの対戦で協調が成立した時刻の分布を表示したものである. 一見して分布は大きく 2 つのピークに分かれることがわかる. 1 つは, 比較的短時間 (10-1000) に協調が成立するもの, もう 1 つは協調までにかなりの時間 ($> 10^5$) を要するものである. 長い時間の方は, たまたま生じた協調シーケンス (G,P,G,P,...) が学習に必要な時間だけ反復することで協調モードに移行する, そのために要する時間であると考えられる. それに対して短い時間の方は, 一方, または両方のプレーヤが意図的に協調を指向したために, 偶然の協調シーケンスが生じるのを待たずに協調を実現できたものである.

人との対戦で一定の協調を誘い出すことができた N 戦略と Q 学習との対戦を見てみると, 分布の大多数は右の山にあり, 少数が左の山になっている. これは, 初期に協調に誘導することに失敗すると, Q 学習は Nash 均衡解を学習してしまい, 協調シーケンスを出しにくくなるからだと思われる. もう一つ興味あるのは FCAQ 戦略の振る舞いである. この戦略は, 相手が Q 学習だと仮定して相手の協調を誘い出す行動をとるように設計されており, ほぼ期待通りに相手を協調に誘導することに成功している. このように, 相手の学習アルゴリズムが推測できれば, 対戦のなかで相手を協調戦略に誘導することは不可能ではない. 一方 Q 学習同士の対戦では, 短時間での協調はほとんど見られないのは, 両者が偶然同時に協調シーケンスを生成するということが起こり得ないからであろう.

図 8 は, SQ1 同士の対戦で協調行動を獲得した「学習後 Q 戦略」をいくつかの戦略と対戦させたものである. N, FCAQ 戦略と対戦させると期待通りに短期

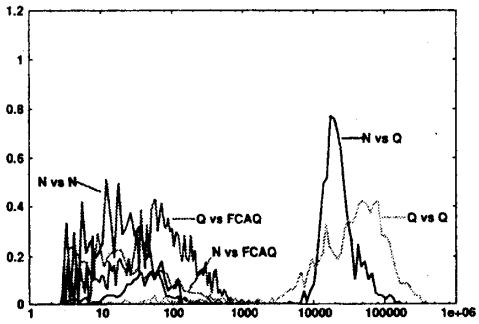

図 7　協調開始時間の分布：N, Q, FCAQ 戦略

間で協調を実現する．学習後 Q 戦略同士の対戦では，協調を再開するまでに少しの時間がかかっている．興味あるのは（未学習）Q 学習戦略との対戦で，せっかく協調を学習したのにもかかわらず，未学習の Q 学習に引きずられて協調実現には Q 学習同士の対戦とほぼ同じ時間を要している．

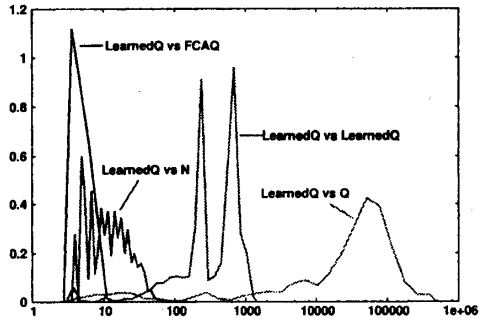

図 8　協調開始時間の分布：学習後 Q 戦略

8. 考察とまとめ

人の場合や学習プログラムのでの経験から，相手の行動を読むことの意味を検討してみよう．

学習プログラムは，(おそらく人の場合も）過去の行動履歴を用いることで協調的行動を実現することができる．しかしながら，制約のない探索でこの解を見つけるにはかなりの対戦を積み重ねる必要がある．これにたいして，一方が一貫して協調に誘導するための行動をとると，短時間で学習エージェントを協調行動に導くことができる．しかしながら，そのためには適切な相手モデルと，一貫した「意思」が必要となる．

我々はこの「短時間で学習エージェントを協調行動に導く」行動を心を読む，と名付けたいと思う．そのアルゴリスムは次のようなものである．

1. これまでの相手の振る舞いから，相手がどのようなエージェントであるかを推定する．これには ϵ を用いて相手を探索することも含まれる．
2. ゲームの構造を解析して，現在の相手に対して自己がもっとも有利となる安定解を発見する．
3. 相手を，上記安定解へ誘導する戦略を実行する．

1. で重要なことは相手が学習エージェントかどうかの判断である．学習エージェントでなければ，3. で相手を誘導することはできない．学習エージェントと分かれば，k 次マルコフ学習戦略と仮定することで相手をモデル化できる．

しかしながら，ここで気をつけなければならないことは，相手も同じようにこちらの心を読んでいる可能性があるということである．一つの対策は，再帰を含んだ相手モデル[7] を作ることである．しかしながら，これは自己が相手に読み勝てることを必要とする．そうでなければ，相手を誘導するつもりが逆に相手に誘導されることにもなりかねない．もっと安直な解決策は，2. を「自己がもっとも有利になる」ではなく，「双方が納得できる」という方向に修正することである．目指すゴールが相手にとっても有利なものであれば，相手に読み勝てなくても，いや相手の誘導に乗ってしまっても自己の目的を達することができる．

残念ながら，我々はこれらの戦略を計算機上で実現できたという訳ではない．現在は，それぞれの部分を人手で解析して，その妥当性を確認している状況である．たとえば，ここでは述べるゆとりはないが，我々はすでに対戦履歴から相手が学習エージェントであるかどうかや，相手の使用している履歴長などを推定するアルゴリズムを開発，検証しているところである．

参 考 文 献

1) Axelrod,R.: *The Evolution of Cooperation*, Basic Books Inc., (1984), 松田裕之訳, つきあい方の科学,HBJ 出版局,(1987).
2) Poundstone, W., *Prisoner's Dilemma*, Doubleday (1992), 松浦俊輔他訳, 囚人のジレンマ, 青土社 (1995).
3) 山岸俊男：社会的ジレンマのしくみ, サイエンス社 (1995).
4) 伊藤 昭：「心を読む能力の創発 —マルチプレイヤー囚人のジレンマゲーム」, 認知科学, Vol.6 No.2, (1999).
5) 鈴木光男, 新ゲーム理論, 勁草書房 (1994).
6) Y. Shoham, R. Powers, and T. Grenager, "On the Agenda(s) of Research on Multi-Agent Learning," 2004 AAAI Fall Symposium on Artificial Multi-Agent Learning (2004).
7) 高野雅典、加藤正浩、有田隆也：「心の理論における再帰のレベルの進化に関する構成論的手法に基づく検討」, 認知科学, Vol.12, No.3, pp.221-233 (2005).