

マルチエージェント強化学習による交渉問題へのアプローチ

伊藤 昭[†] 水野 将史[†] 松本 達明[†] 寺田 和憲[†]

[†] 岐阜大学工学部 〒501-1193 岐阜市柳戸 1-1

E-mail: †ai@info.gifu-u.ac.jp

あらまし 交渉問題とは、それぞれ独立に最善を尽くすよりも話し合って協力して行動した方が有利なゲーム理論的状况で、どのようにして協力解を見いだせば良いかという問題である。我々はこれまで提案されているような「万人が認める規範解」を理論的に検討するのではなく、各エージェントが自己の利益を追求する中で、双方が自己にとって最良の妥協点を求める動力学の結果として協力解が求まるものと考え、これを記述する手法として、各エージェントは相手の行動を観察し、得られた情報に基づき自己の行動を調節する学習エージェントであると考え、マルチエージェント強化学習の理論を適用する。本手法により、交渉成立の条件や妥協点の決定の機構について新しい視点からの見方を提供する。

キーワード 交渉問題, マルチエージェント, 強化学習 履歴を用いたQ学習

An Approach to Bargaining Problem Through Multi-Agent Reinforcement Learning

Akira ITO[†], Masafumi MIZUNO[†], Tatsuaki MATSUMOTO[†], and Kazunori TERADA[†]

[†] Faculty of Engineering, Gifu University Yanagido 1-1, Gifu, 501-1193 Japan

E-mail: †ai@info.gifu-u.ac.jp

Abstract Bargaining problem is how to find the contract point where the cooperation is preferable, but we must fight for the advantageous agreement. We approach bargaining problem from the multi agent learning standpoint. Each agent observes the opponent's behavior and tries to increase the expected interests by adjusting its own action appropriately. The cooperative solution is obtained through the pursuit of the self-interests of each agent. This method propose a new view for the meaning of the cooperative solution in bargaining problem.

Key words bargaining problem, multi agent, reinforcement learning, Q learning using history

1. はじめに

交渉問題とは、それぞれ独立に最善を尽くすよりも話し合って協力して行動した方が有利なゲーム理論的状况で、どのようにして協力解(妥協点)を見いだせば良いかという問題である[1]。たとえば、隣同士の二人が毎日町まで荷物を運ばねばならないとすると、交代で隣の人の荷物を運んであげればお互いにメリットがあるであろう。このように状況が対称的な場合には、自然な協力解を見いだすことは比較的簡単である。

しかしながら状況が非対称なときには、妥協点を見いだすのはそう簡単ではない。たとえば、一人の人は毎日手紙を一通、もう一人の人は重い荷物を運ばねばならないとすれば、どのように分担すれば良いのだろうか。重い荷物を町まで運ばねばならない人の方が相対的に多くの回数を負担すべきだという気もする。客観的に二人の「仕事量」を比較できれば良いという考え

方もあるが、一方の人は他方よりも「町の雰囲気を楽しみたい」と考えていて、町に出かけることを相対的に苦痛に感じないかも知れない。このように様々な要素を考慮にいと、妥協点はどうなるのだろうか。

この問題は交渉問題(bargaining problem)と呼ばれて、経済学、ゲーム理論ではNashによる提案[2]以来、様々な議論されてきた。その考え方の基本は以下のように要約される。

1. この問題には、このままでは科学的な「正解」というものは存在し得ない。
2. したがって、解が満たすべきいくつかの公準(Axiom)を定め、それを満たす解を求める。この公準が万人が認める妥当なものであれば、その結果導かれる解も万人に認められるであろう。
3. しかしながら、多くの人が予想するように、万人が一致して認める公準など存在せず、様々な公準が、またそこから導かれる様々な解が提案されてきた。

我々はこのような規範的接近を採るのではなく、二個のエージェントがインタラクトする中で、双方が自己にとって最良の妥協点を求める動力学 (dynamics) の結果として解が求まるものとする。またこの動力学を記述する手法として、各エージェントは相手の行動を観察し、得られた情報に基づき自己の行動を調節する学習エージェントであると考え、マルチエージェント強化学習の理論 [3] を適用する。

すでに述べたように、対称的な問題では公平性 (fairness) の観点から、容易に自然な妥協点が見いだされる。もちろん、公平な妥協点が動力学のモデルで本当に実現されるのか、またそれはどのような過程によるのかについては、別途検討を必要とする問題である。我々は、これについても「1・2・5じゃんけん」を用いてマルチエージェント学習の立場から分析を行っているが、本論文では分析の対称を非対称ゲームに限ることにする。

2. 交渉問題

我々が取り上げるのは、1. で述べたような、「協力すればお互いに利益があるが、その妥協点をめぐって双方が争わねばならない」問題である。このような問題は、ゲーム理論では非零和二人ゲームを用いて定式化できる。話を具体的 (数学的) にするため、表 1 で与えられる非零和二人ゲームを考える。

	1	2
1	8,4	2,3
2	6,2	4,6

表 1 交渉問題の利得行列

縦はプレーヤ P1 の戦略 $s_1 \in \{1, 2\}$, 横はプレーヤ P2 の戦略 $s_2 \in \{1, 2\}$ である。また表中の値 r_1, r_2 は、それぞれの戦略が採られたときの P1, P2 の利得である。なお、ゲーム理論では利得について次のことが想定されている。

1. 利得はそれぞれのプレーヤにとっては加算的である、すなわち、利得 2 を 2 回得ることは、利得 4 を 1 回得ることと同じ程度に好ましい。
2. 二人のプレーヤの利得を直接比較することは意味がない。すなわち、相手の得る利得 2 の好ましさが、自分の得る利得 2 の好ましさと同程度であると考えられる根拠はない。

一番目の仮定はゲーム理論の根幹であり、ほぼ全ての研究者により認められている。二番目の仮定については様々な議論があるところである。

さて、このゲームの Nash 均衡点とそのときの利得は

$$(s_1, s_2)_A = (1, 1), (r_1, r_2)_A = (8, 4)$$

$$(s_1, s_2)_B = (2, 2), (r_1, r_2)_B = (4, 6)$$

の A, B 二つであり、ともにパレート (Pareto) 最適 [1] である。また、プレーヤ P1, P2 それぞれの最低保証点 (相手がどのように行動しても確保できる得点) (c_1, c_2) は

$$c_1 = \max_i \min_j a_{ij} = 4, c_2 = \max_j \min_i b_{ij} = 3$$

で与えられる。

次に、この問題の混合戦略での均衡解を求めてみる。いま、P1 が $s_1 = 1$ を確率 p で、 $s_1 = 2$ を確率 $(1-p)$ で採り、また P2

が $s_2 = 1$ を確率 q で、 $s_2 = 2$ を確率 $(1-q)$ で採るとする。このとき、P1, P2 の利得の期待値 (f_1, f_2) は表 1 の利得を用いると、

$$f_1 = 4 + 2q + (4q - 2)p, f_2 = 6 - 3p + (5p - 4)q$$

と求められる。したがって、P1 の P2 の戦略に対する最適反応戦略は

$$p = 0 (q < 1/2), [0, 1] (q = 1/2), 1 (q > 1/2),$$

となる。同じく、P2 の P1 の戦略に対する最適反応戦略は

$$q = 0 (p < 4/5), [0, 1] (p = 4/5), 1 (p > 4/5)$$

となる。Nash 均衡点はこれらの共通部分として

$$(p, q)_A = (0, 0), (p, q)_B = (1, 1), (p, q)_M = (4/5, 1/2)$$

の 3 個となる。ちなみに、 $(p, q)_M$ での利得は $(r_1, r_2)_M = (5, 18/5)$ となる。

人がこのようなゲームを繰り返し対戦しなければならないとしたら、どのような戦略で望めば良いのだろうか。混合戦略の範囲で利得 (f_1, f_2) の実現可能領域を求めると、図 1 の斜線領域となる。M は混合戦略の Nash 均衡点である。中心部の曲線は $F(f_1, f_2) = 25f_1^2 - 40f_1f_2 + 16f_2^2 - 132f_1 + 64f_2 + 324 = 0$ で決まる楕円の一部分である。双方にとってより有利に見える線分 AB 上の点は、それぞれが独立に一定の確率で戦略を選択する混合戦略の範囲では実現できない。

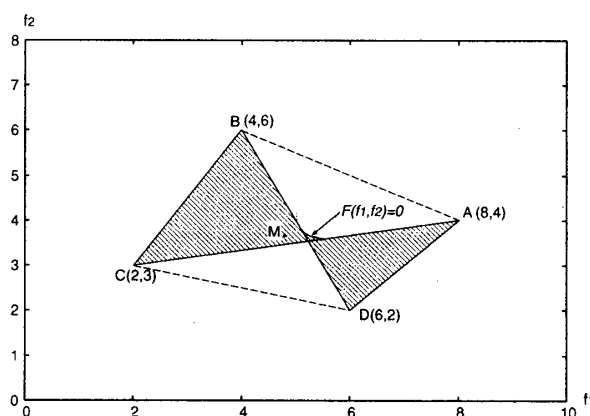


図 1 交渉問題 (表 1) での平均報酬の実現可能領域

一方、無限回繰り返しゲームと考えたときの、平均利得の実現可能領域は図 1 の四角形内部となる。実際点 A, B, C, D の任意の 2 点を結ぶ直線上の点は、それに対応する戦略を適当な割合でプレーすることで実現できる。たとえば、線分 AB を x 対 $(1-x)$ に内分する点 X は、協調して $(s_1, s_2)_A = (1, 1)$ を x , $(s_1, s_2)_B = (2, 2)$ を $(1-x)$ の割合でプレーすることで実現可能である。明らかに、繰り返しゲームでは直線 AB 上の点のみがパレート最適であり、そこでは平均報酬 (r_1, r_2) は関係 $r_1 + 2r_2 = 16$ を満している。混合戦略の Nash 均衡点 M (を連続してプレーすること) は、繰り返しゲームにおいてはパレート最適ではない。

交渉が可能であれば、それぞれが独立に (最適な) 混合戦略をとるよりも、協力して A または B をプレーする方が双方にとって利益があり、協調する動機は十分にある。しかしながら協調が成立するためには、妥協点 X (または A をプレーする割合 x)

を交渉 (bargain) により合意する必要がある。これが交渉問題と呼ばれるものである。

3. 交渉の妥協点

2. で述べた交渉問題について、交渉の妥協点 X を決める合理的な理論はあるのであろうか。無限回繰り返しゲームについてのフォーク定理によれば、多角形内で線分 AB の任意の近傍には、その点を平均利得とする「サブゲーム完全な Nash 均衡」解が存在することが知られている。しかしながら、これは線分 AB 上の任意の点が、無限回繰り返しゲームでは実現可能と主張しているだけであり、その中のどの点が「望ましい」のかについては何も教えてくれない。

鈴木 [1] によると、交渉問題での Nash 解 (Nash が提案した交渉問題での妥協点=協利解のことで、Nash 均衡とは別の概念である) は A を $3/4$ 、 B を $1/4$ でプレイすることで、その平均利得は $(r_1, r_2)_N = (7, 4.5)$ であるという。この解は、2. で述べた最低保証点 $(c_1, c_2) = (4, 3)$ を基準点とし、そこからの増分の積を最大化することで得られる。すなわち、妥協点 X での得点の平均値を (r_1, r_2) として、 $f_u = (r_1 - c_1)(r_2 - c_2)$ を最大化するように X を決めるというものである。利得の増分の積を最大化することは、利得行列の線形変換に対して協利解が不変という要請から導かれる。線形変換不変性は、前に述べた「利得は個々のプレーヤーにとっては加算的であるが、プレーヤー間の利得を直接比較することはできない」ということに対応している。

しかしながら鈴木も述べているように、この解には「いくつかの公準を満たす無矛盾な解」という以上に深い意味はなく、そもそも「正しい解」というものは数学的には存在しない。おそらくこの解について一番意見の分れるところは、増分の積を最大化することではなく、基準点、すなわち「交渉が失敗だったとき自分が得るであろうと思われる利得の期待値」をどこに見做すかということにあると思われる。

MiniMax 原理から求められる最低保証点を基準点とする根拠はそれほど強くない。たとえば、お互いにランダムに行動したときの得点の期待値 $(r_1, r_2)_R = (5, 3.75)$ も一つの候補として考えられるであろう。より自然なのは、双方が自己の主張を通そうとしたときの均衡点 $(s_1, s_2) = (1, 2)$ での利得 $(r_1, r_2) = (2, 3)$ である。しかしながら、この範囲の議論ではそれぞれの主張は「もっともらしい」とは言えても、それ以上の根拠を与えることはできない。

より説得力がある議論として、鈴木は「脅しの戦略」を紹介している [4]。それは、本来自分にとっては意味のない相手の利得を相手を制御するために利用しようというもので、一旦交渉問題だとの共通認識が得られれば、妥協点を決める問題を相手と自分の取り分をあらそう零和ゲームと見做して争おうというのである。具体的には、プレーヤ 1, 2 の利得行列を R_1, R_2 として、線分 AB 上では利得ベクトル (r_1, r_2) が $r_1 + 2r_2 = 16$ を満たすことを考えると、プレーヤーは $R_1 - 2R_2$ を、プレーヤ 2 は $2R_1 - R_1$ 利得行列のように考えて行動することになり、Nash 均衡解としては $(s_1, s_2) = (1, 2)$ を得る。

いずれにせよ、基準点の利得 (c_1, c_2) が求まると、Nash 解

は $r_1 + 2r_2 = 16$ の条件の下で $f_u = (r_1 - c_1)(r_2 - c_2)$ を最大化する (r_1, r_2) として求まり、それぞれの平均利得は $r_1 = 8 - 4y, r_2 = 4 + 2y, y = (2c_2 - c_1)/8$ で与えられる。ここで、 y は B をプレイする割合 (A をプレイする割合 $x = 1 - y$) である。さまざまな基準点について Nash 解をまとめると表 2 のようになる。

	基準点	Nash 解	A の割合 x
MinMax	(4, 3)	(7, 4.5)	0.75
Random	(5, 3.75)	(6.75, 4.625)	0.6875
脅し	(2, 3)	(6, 5)	0.5

表 2 交渉問題の様々な基準点

4. 履歴を用いた強化学習

我々のアプローチは、実際に強化学習をするプログラムを対戦させることでどのようにして交渉が行われるのかを観察し、そこから解の意味を考え直そうというものである。学習エージェントは、これまでの対戦履歴を使って自己の報酬の期待値を最大化するように政策 (戦略) を決定する。以下では、履歴を用いた強化学習を Q 学習の枠組みを用いて定式化する。交渉ゲームにおける学習の目標は、表 1 の利得行列で定まるゲームを無限 (不定) 回繰り返し、その過程で自己の平均得点を最大化するような戦略 (政策) を見つけることである。

以下では繰り返し非零和二人ゲームを想定して、学習エージェントが共存する状況下での 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。これを学習として見るときには、手を行動、利得を報酬、無限回繰り返しゲームの戦略を政策と呼ぶ。また、学習アルゴリズムの中では自己を m 、相手を o として区別する。

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、時間普遍的な形で過去 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

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

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

$$a^m(t-2), a^o(t-2), \dots, a^m(t-k), a^o(t-k)\}$$

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

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

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過程としてモデル化できる。

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（あらかじめ回数 k を固定しておいた上で）ここで述べた意味での学習政策に限るものとする。

今相手の政策を固定して考える。自己の政策を π として、 $Q^\pi(S, a^m, a^o)$ を、状態 S で自分が手 a^m を、相手が a^o を選んだときの得られる γ で割引された報酬の期待値とする。すなわち、 a^m, a^o によって S_t から遷移する状態を S_{t+1} 、そこから政策 π （とあらかじめ固定して考えている相手の政策）により生成される状態、行動系列を $\{S_\tau, a^m(\tau), a^o(\tau)\}, \tau = t+1, t+2, \dots$ 、として

$$Q^\pi(S, a^m, a^o) = r(a^m, a^o) + \sum_{\tau=t+1}^{\infty} \gamma^{\tau-t} r(a^m(\tau), a^o(\tau))$$

とする。ここで $r(a^m, a^o)$ は手の組 (a^m, a^o) で得られる報酬である。

ここで、自己の政策 π が、相手の政策を固定して考えたときの最強の（＝ Q 値を最大化する）政策であるとする、

$$Q^\pi(S, a^m, a^o) = r(a^m, a^o) + \gamma \max_{a^m} \bar{Q}^\pi(S', a^m)$$

$$\bar{Q}^\pi(S', a^m) = \sum_{a^o} p(a^o|S', a^m) Q^\pi(S', a^m, a^o)$$

の関係を得る。ここで S' は、 S から行動 (a^m, a^o) によって遷移した状態、 $p(a^o|S', a^m)$ は状態 S' で自分が行動 a^m をとったときの相手の行動 a^o の確率である。また、 Q 学習に倣って逐次近似法で Q の値を求めることにすると、次の式を得る。

$$Q(S, a^m, a^o) \leftarrow (1 - \alpha) Q(S, a^m, a^o) + \alpha(r(a^m, a^o) + \gamma \max_{a^m} \bar{Q}(S', a^m))$$

$$\bar{Q}(S', a^m) = \sum_{a^o} p(a^o|S', a^m) Q(S', a^m, a^o)$$

相手が、過去 k 回の履歴を用いて次の手を決める固定戦略 (k 次マルコフ戦略と呼ぶ) と仮定すると、過去 k 回の対戦履歴を状態とすることで、相手を含めた外界がマルコフ決定過程 (MDP) としてモデル化できる。この時は $p(a^o|S', a^m)$ は a^m に依存しない。 a^m への依存性は、これから自分のとる行動により相手の行動の生起確率が変化することを意味し、相手がこちらの手を「読む」能力を想定したことになる。

$p(a^o|S, a^m)$ は状態 S での行動の生起確率を観測することで求められる。具体的には、時刻 t で状態 S_t にあり、行動 $(a^m(t), a^o(t))$ が観測されたとすると、

$$p(a^m, a^o|S_t) \leftarrow p(a^m, a^o|S_t) + \delta$$

$$(a^m = a^m(t), a^o = a^o(t) \text{ のとき})$$

$$p(a^m, a^o|S_t) \leftarrow (1 - \delta) p(a^m, a^o|S_t)$$

$$p(a^o|S_t, a^m) = p(a^m, a^o|S_t) / (\sum_{a^o} p(a^m, a^o|S_t))$$

により確率が更新される。

行動選択は ϵ -greedy と Boltzmann 型とを組み合わせで行う。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

$p(a) = C \exp(\bar{Q}(S, a)/T)$ の確率で $a \in A^m$ を選択

なお、ここで ϵ は 1 より小さな数、 T は探索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 k を用いて PQk (履歴長 k 確率予測 Q 学習) と表す。

5. 実験結果 1—履歴長の役割

我々は、履歴を用いた確率予測 Q 学習エージェント PQk を履歴長をさまざまに変えて対戦させてみた。その結果をいくつかの指標を用いて説明する。P1, P2 の平均得点を r_1, r_2 とすると、 $s = r_1 + 2r_2$ が交渉の成立の程度を表す。 s は直線 AB 上でのみ 16 となり、それが実現できるのは、双方が協調して A または B を採るときのみである。もう一つの指標は、 $x = BX/AB$ で、妥協点 X が AB 上のどこにあるかを表す。P1 の有利な A 点では $x = 1$ 、また P2 の有利な B 点では $x = 0$ である。

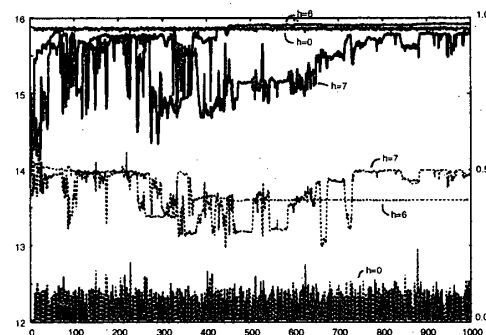


図 2 PQk-PQk 対戦における平均得点の時間的変化

まずは、同じ履歴長同士 (PQk-PQk) の対戦を行って見た。実験は、表 1 で与えられる利得を用いて、それぞれの政策を 10^8 回繰り返し対戦させた。実験に使用したパラメタは、 $\alpha = 0.1, \gamma = 0.9, \delta = 0.1, \epsilon = 0.01, T = 0.4(P1), T = 0.2(P2)$ である。 T の値が P1, P2 で異なるのは、交渉のポイントでは双方の利得が $r_1 + 2r_2 = 16$ を満たしながら争われるため、P2 の利得 1 が P1 の利得 2 に交換可能と見立ててのことである。

そのときの s, x の時間的振る舞いを、履歴長 $h = 0, 6, 7$ について図 2 に示す。グラフの上方に張り付いている実線が s であり、値は左目盛を用いる。グラフの下方の波線が x であり、値は右目盛を用いる。横軸は対戦回数で 10^5 ごとの平均を示している。履歴長 7 を除いて s は急速に上限値に収束しており、早い時間で協調行動が発現していることがわかる。妥協点での x は履歴長 6, 7 では 0.5 に近いところにあるが、履歴長 0 では、0.1 の近くを激しく変動していることがわかる。

つぎに、 s, x の漸近的 (収束後の) 振る舞いを見るため、上記の 10^8 回対戦を 1 ラウンドとし、乱数を変えて 10 ラウンドの実験を行った。漸近値としては各ラウンド 10^8 の内の後半の 5×10^7 を採ることにした。これは、図 2 によれば、履歴長 7 を除いて s が前半の 5×10^7 までではほぼ上限値に収束していることから、後半では漸近的振る舞いと見做し得ると考えてのことである。

履歴長を変えたときの 10 ラウンドの妥協点 x の分布を図 3

に示す。横軸が x 、縦軸はその値がとられる頻度である。また平均値を矢印で示してある。図を見ると多くの場合 x の収束値は簡単な分数 n/m になっており、A, B が $n : (m - n)$ の割合で交替して採られていることが理解される。なお、履歴長を h とすると周期 $(h + 1)$ を超えるパターンを学習することはできないため、可能な収束値の分母 m は $m \leq (h + 1)$ に制限される。

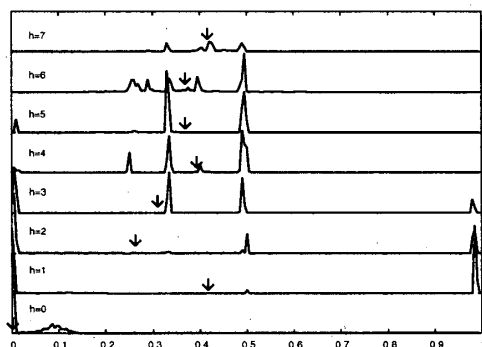


図3 PQk-PQk 対戦における妥協点 x の分布

同じデータを用いて s, x および x の標準偏差 $\sigma(x)$ について 10 ラウンドの平均を求めた結果を表 3 に示す。その結果を要約すると以下の通りである。

1. 履歴長が大きくなるにしたがい、 x の平均値が大きくなり、 x の分散が小さくなる。
2. $h \leq 6$ では、 s は上限値に近く、協調的行動がうまく行われている。
3. $h = 7$ では、収束は遅く s が上限値 16 から大きく落ち込んでしまう。
3. は、履歴長 7 では、協調することが与えられた時間内に実現できなかったことを意味する。このことは図 2 によっても確認できる。

	PQ0	PQ1	PQ2	PQ3	PQ4	PQ5	PQ6	PQ7
s	15.85	15.87	15.88	15.88	15.87	15.85	15.72	15.55
x	0.039	0.417	0.269	0.305	0.396	0.372	0.373	0.412
$\sigma(x)$	0.045	0.473	0.393	0.299	0.123	0.131	0.095	0.057

表3 PQk-PQk 対戦における $s, x, \sigma(x)$

PQ6-PQk 対戦

	PQ0	PQ1	PQ2	PQ3	PQ4	PQ5
s	15.88	15.87	15.87	15.81	15.86	15.80
x	0.004	0.004	0.038	0.193	0.227	0.389
$\sigma(x)$	0.000	0.001	0.099	0.191	0.239	0.119

PQk-PQ6 対戦

	PQ0	PQ1	PQ2	PQ3	PQ4	PQ5
s	15.80	15.84	15.87	15.78	15.77	15.83
x	0.990	0.695	0.434	0.543	0.447	0.450
$\sigma(x)$	0.003	0.244	0.198	0.241	0.077	0.085

表4 PQ6-PQk, PQk-PQ6 対戦における $s, x, \sigma(x)$

次に、収束が見られた最大の履歴長 $h = 6$ の PQ6 とさまざ

まな履歴長 k の PQk とを対戦させたときの漸近的振る舞いを表 4 に示す。PQk 同士の対戦と同じく、乱数を変えて 10 ラウンド行い、1 ラウンド 10^8 回の対戦のうち、後半の 5×10^7 を漸近値として用いている。政策の組み合わせは、上段は P1 を PQ6 とし、P2 を PQk としたもの、下段はその逆である。また、表から妥協点での x のみを取り出してグラフにしたものを図 4 に示す。次のことが表や図から読み取れる。

- 1) 全ての組み合わせで、 s は上限近くに収束しており、協調が成立していることがわかる。
- 2) 履歴長の大きいものと、小さいものが対戦すると、小さい方が有利な立場に立てるようにみえる。

2) は、履歴長が長い方が「適応能力が大きい」と考えると、交渉問題では適応能力の大きい方がどうしても先に譲歩をしてしまうのではないと思われる。これは、ゲーム理論で良く知られているタカ・ハトゲーム [5] と同じで、タカが有利と言うわけではない。実際、履歴長が小さなもの同士が対戦すると、妥協点は両端に偏り、「一か八か」の戦いとなる。それに対して、より大きな履歴長では平均得点の分散は小さくなり、お互いにとって妥協点を見いだすことが可能となる。

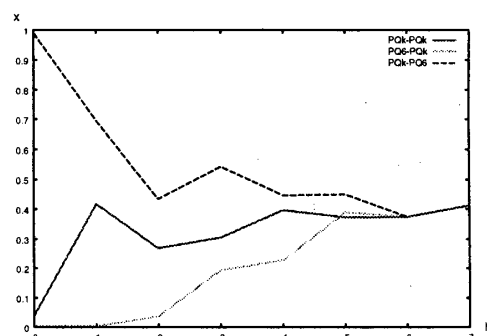


図4 PQk-PQk 対戦における妥協点 x

6. 実験結果 2 — 利得行列修正の効果

これまでの履歴長が妥協点の決定にどのような役割を果たすのかを見てきた。では、履歴長を固定（たとえば $h = 6$ ）したとき、どのようにして妥協点は決まるのであろうか。とくに、利得行列の修正は妥協点の決定にどのような役割を果たすのだろうか。これを見るために、戦略の組 $(s_1, s_2) = (2, 1)$ に対する利得 $(r_1, r_2) = (6, 2)$ の値を変えて、その振る舞いを調べてみる。一見 A, B での利得を変更する方が本質的であるようだが、実は「利得の線形変換によっては、ゲームの定性的性質は変更されない」ということから、A, B における利得の変更は、他の非対角要素の利得の変更と同等であることがわかる。さらに、3. で述べた様々な基準点では戦略の組 $(s_1, s_2) = (1, 2)$ (利得行列の右上) の値は使われていたが、戦略の組 $(s_1, s_2) = (2, 1)$ での値は使われていなかった。我々の目的は、このような非対角要素の利得の値が妥協点の決定にどのような役割をもつのかを実験的に調べることである。

まず、 r_2 を変えた結果を表 5 に示す。興味深いことに r_2 を大きくすると、本来それは P2 にとって有利に働くべきものが、妥

協点を P1 の有利な方向に移動させるのである。これは、どのように解釈しても基準点を用いて Nash 解を求める枠組みからは予想できないことである。しかしながら、強化学習のアルゴリズムを考えると、 $s_2 = 1$ に対する P2 の得点を高めることで、P2 の $s_2 = 1$ に対する拒否反応を相対的に弱めて、結果として P1 に有利な A 点の出現確率をあげているのではないかと推測される。

	(6, 0)	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)
s	15.73	15.69	15.72	15.82	15.82	15.83
x	0.376	0.364	0.373	0.466	0.503	0.489
$\sigma(x)$	0.092	0.168	0.095	0.049	0.054	0.078

表 5 利得 r_2 を変えたときの $s, x, \sigma(x)$ の振る舞い

	(-2, 2)	(0, 2)	(2, 2)	(4, 2)	(5, 2)	(6, 2)
s	15.82	15.83	15.79	15.76	15.76	15.72
x	0.555	0.531	0.489	0.431	0.447	0.373
$\sigma(x)$	0.079	0.053	0.063	0.065	0.094	0.095

表 6 利得 r_1 を変えたときの $s, x, \sigma(x)$ の振る舞い

同じく r_1 を変えた結果を表 6 に示す。やはり、 r_1 を小さくすると、妥協点の x は P1 に有利な方向に移動する。

7. 考 察

交渉問題を協力ゲームとみる従来の立場から見れば、我々が扱っている問題は交渉問題ではない。プレーヤ間には何も交渉が行われていないからである。しかしながら、交渉問題ではどのようにして妥協点が決まるのかを考えると、そこには「交渉不成立」のときの双方の損得の情報が必要となる。人は、規範解を考えると過去（類似の）ゲームの経験をもとに、どのようなことが生じるかを心の中でシミュレーションし、妥協することの損得を計算する。我々がここで行った計算はそのような心のシミュレーションに相当するものと考えている。

我々が得た興味深い結果の一つは、交渉問題を解決するためには必ずしも「交渉」しなければならないという訳ではないということである。履歴を用いて各エージェントが最善の行動を模索するなかで、自然に協調的行動が発現し得るのである。これは、交渉問題とは異なる利得行列を用いる囚人のジレンマゲームでは経験的に知られていることである。

しかしながら、実験結果はこれまでの理論と少し異なっている。表 2 にあるように、これまでに提案された様々な解は $x \geq 0.5$ を提案しているのに対して、実験結果は $x < 0.5$ を示唆している。我々は 6 において、非対角要素の利得にどのように影響されるかを調べた。その結果、非対角要素の利得を自己に不利に修正することが、結果として妥協点を自己に有利な位置に移動させ得ることを示した。

しかしながら、これが常に成り立つものと考えすることはできない。これは、対戦相手がお互いに学習エージェントであることにその理由がある。背水の陣のような明らかに自己に不利な戦略が、時として有効であることは経験的には良く知られてい

る。しかしながら、退路を立つことが合理的に考えれば有利であるはずはなく、それが有利である状況があるとすれば、それが相手や自己の行動を変更させることにある。同じように、非対角要素の自己の利得を低くすることは、それが自己の行動を変え、さらにそれが相手の行動選択に影響を与えることで、はじめて効果を持つのである。

もちろん、交渉問題とはまさに様々な状況を総合的に利用して判断を行う問題であり、理想的状況で規範解を議論するだけでは解決しないものである、というのが我々の立場である。そのような観点からすれば、適切な相手モデルを用いて相手を誘導することが、まさに交渉問題の本質部分をとらえているものと考えられる。交渉問題はゲーム理論として利得行列を与えられるだけでは正解は存在しないが、具体的な人やエージェントがそれをプレイするときには、「より良い」戦略が存在し得る問題となるのである。

しかしながら、依然として「何が正解か」という問題は解決していない。一つの判断基準は、人がこのゲームを繰り返し対戦せねばならないとき、

1. 話し合いをしなくても協調が達成できるのか、
2. 話し合いを許したとき、結局どのような妥協点が達成されるのか、

を実験的に調べてみることである。もちろん、話し合いをしても、何らかの制約をかけない限り話し合いはまとまるはずはない。たとえば、時間制限をもうけて、話し合いの成立の如何に関わらず、ゲームを開始する。そのとき、人は交渉不成立の危険に直面するが、各自は交渉不成立の確率、不成立の場合の損害を計算して、妥協点を決めることになる。果たしてこれまで提案されてきた規範解に近い結果が得られるのであろうか、またはここで行ったような強化学習のシミュレーションに近い結果が得られるのであろうか、興味深い研究課題である。

もう一つは、このようなゲームを囚人のジレンマゲームの時のように、プログラムコンテストとして行うことである。しかしながら囚人のジレンマゲームとの大きな違いは、囚人のジレンマゲームの時に存在した「上品な」戦略というものは、ここでは存在し得ないことである。単純に協調的だけでは決して高い得点は稼ぐことができない。相手のモデルを作り、相手の行動予測に基づいて最適な行動を生成することが求められるのである [6]。このような状況でどのような戦略（政策）が高得点をあげ得るのであろうか。これも興味深い研究課題である。

文 献

- [1] 鈴木光男：「新ゲーム理論」、勁草書房、1994。
- [2] Nash, J.: "The bargaining problem," *Econometrica*, 18, pp.155-162, 1950.
- [3] Sutton, R.S and Barto, A.G., *Reinforcement learning, an introduction*, A Bradford Book, The MIT Press(1998).
- [4] 鈴木光男：「ゲーム理論入門」、共立出版、1981。
- [5] Smith, J. M.: "Evolution and the Theory of Games", Cambridge Univ. Press, 1982, 寺本英, 梯正之訳、「進化とゲーム理論-闘争の論理」、産業図書、1985。
- [6] 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。