履歴を用いたQ学習による交渉問題へのアプローチ

水野将史 松本達明 伊藤昭 寺田和憲
岐阜大学工学部

1 はじめに

交渉問題は、それぞれのエージェントが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況である。この問題は、経済学・ゲーム理論ではNashによる提案以来[1]、妥当な協力解の満たすべき条件をめぐって、様々に議論され続けてきた。

我々は、交渉問題にこれまでのような規範的接近をとるのではなく、2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(dynamic)の結果として解が求まるものと考える。この動力学を記述する方法として、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。

2 交渉問題

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

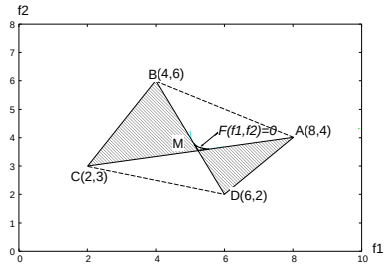

図 1: 交渉問題の利得行列と現実可能領域

我々が取り上げるのは、図1にある利得表の非零和二人ゲームであり、「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なければならない」問題である[2]。この問題のNash均衡点は図1のA,B二つであり、その時の利得は(8,4)(4,6)、ともにパレート(Pareto)最適である。また、それぞれが独立にp, qの確率で1,2を選択する混合戦略をとると、その実現可能領域は図1のようになる。

しかし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、協力して図のAまたはBを一定の割合でプレーする方が双方にとって有益であり、その時の報酬(r_1, r_2)は$r_1 + 2r_2 = 16$(線分AB上)を満たす。しかしながら、協調が成立するためには、妥協点X(Aをプレーする割合x)を交渉によって合意する必要がある。Nashが提案する交渉問題の妥協点＝協力解は、基準点の利得(c_1, c_2)から$r_1 + 2r_2 = 16$の条件の下で$f_u = (r_1 - c_1)(r_2 - c_2)$を最大化する$(r_1, r_2)$である。この外にも基準点には様々な提案があるが、どれも無矛盾な解以上に深い意味を持たない。

3 履歴を使ったQ学習

我々のアプローチは、実際に強化学習エージェント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報酬の期待値を最大化する自己の政策(戦略)を学習する。

まず状態S_tを、自己および相手の過去h手の情報を用いて$S_t = \{a^m(\tau), a^o(\tau)\}_{t-1}^{t-h}$で定義する。この時の自己の政策πとし、$Q^\pi(S, a^m, a^o)$は、状態$S$で自分が手$a_m$、相手が$a_o$を選んだ時の得られるγで割引された報酬の期待値となる。ここで期待値が最大となる政策πが最強である。すなわち、

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

となる。(S'は、Sから行動(a^m, a^o)によって遷移した状態である。)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。

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

また行動選択確率は

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

で更新される。

4 実験/結果

我々は履歴を用いた確率予測Q学習エージェントPQhを履歴長を様々に変えて対戦させてみた。プレイヤー1,2の平均利得をr_1, r_2とすると$s = r_1 + 2r_2$が16に近い程、交渉の精度が高いと言える。また、$x = BX/AB$で妥協点Xの線分AB上の位置を表す。

まずは、同じ履歴長同士(PQh-PQh)の対戦を行った。実験は利得表を用いた対戦を10^8回行うのを1ラウンドとし、乱数を変えて10ラウンド行った。そのときのs, xの時間的振舞い(10ラウンドのうちの1つのラウンド)

[1] An approach to bargaining problem using Q-Learnig with history
Masafumi Mizuno,Tatsuaki Matumoto,
Akira Ito,Kazunori Terada
Faculty of Engineering,Gifu University

を図 2 に示す。図からわかるようにように s は急速に上限近くに収束し、多くの場合 x も一定値に収束する。そこで各ラウンド 10^8 の内、後ろの半分についての平均を採って収束値とした。用いたパラメタは、$\alpha = 0.1$, $\gamma = 0.9$, $\delta = 0.1$, $\varepsilon = 0.01, T = 0.4(\mathrm{P1}), T = 0.2(\mathrm{P2})$ である。T の値が P1 と P2 で異なっているのは、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である。

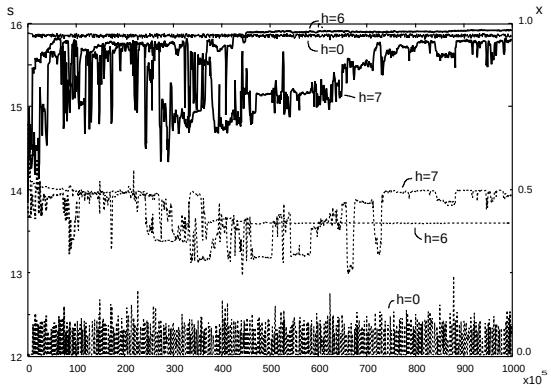

図 2: PQh-PQh 対戦における s, x の時間変化

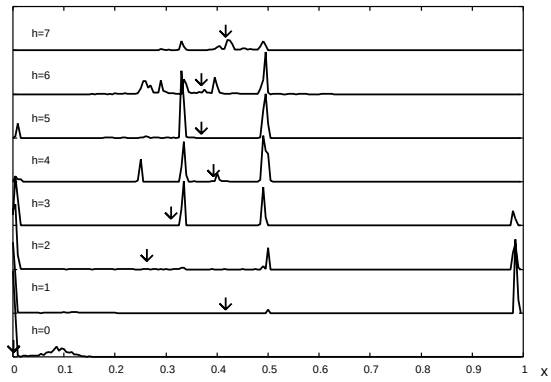

図 3: 履歴長別 x の分布

履歴長を変えたときの、10 ラウンドの x の分布を図 3 に示す。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 $n : m$ の割合で採られる政策であることがわかる。履歴長を h とすると周期 $(h+1)$ を越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。また、履歴長が小さい時ほど x の収束は端に偏り、h の増大に伴い中心近くに移動する。

さらに、同じデータを用いて s, x の収束値の平均をもとめた結果を表 1 に示す。履歴長が大きくなるにしたがって x の平均値が大きくなり、分散が小さくなっている。履歴長 $h = 6$ 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といえる。しかし、$h = 7$ では逆に落ち込んでいるがこれは協調することがこの時間内に学習できなかったことを意味している。

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

表 1: PQh-PQh 対戦における s、 x、$\sigma(x)$

つぎに、収束が見られた最大履歴長 $h = 6$ と様々な履歴長とを対戦させた結果を含め x の平均値を図 3 に示す。PQ6-PQh は P1 の履歴長を 6 に、PQh-PQ6 は P2 の履歴長を 6 に固定したものである。履歴長の大きいものと小さいものが対戦した場合、小さいものの方が優位な利得を挙げている。これは、履歴長の長い方が「適応能力が大きい」ので、どうしても先に譲歩してしまう為と考えられる。

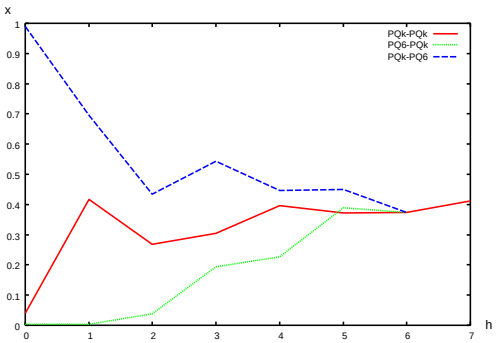

図 4: PQh-PQh,PQ6-PQh,PQh-PQ6 対戦における妥協点 x

5 まとめ

実験結果は、これまでの理論と少し異なっていて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。また、この収束値 x の決定には、$(s_1, s_2) = (2, 1)$ での利得が大きく関与していることが別実験からわかっている。

人がこのような問題に直面したときには、交渉が成立しなかった場合の損得情報を踏まえ過去の経験を下に、心の中でシミュレーションし妥協点を計算する。今回我々が行った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は、ある意味で人が行っているような交渉過程を再現したものと言える。実際、人と人でこの問題に取り組んだ場合、今回のシミュレーションに近い結果が得られるかどうか、興味深い研究課題である。

参考文献

[1] Nash, J.: "The bargaining problem," Econometrica, 18, pp.155-162, 1950.

[2] 鈴木光男：「新ゲーム理論」、勁草書房、1994.