繰り返し対称非零和ゲームの強化学習 —１・２・５じゃんけんを例に—

後藤 強　岩田元志　伊藤 昭　寺田和憲
岐阜大学大学院工学研究科

1 はじめに

我々は, 知的なエージェントがゲーム理論的状況で, どのようにして相手の行動を読み, 自己の最適行動を生成するのかを検討する. そのような問題の一つとして, 非零和ゲームを繰り返し対戦しなければならないとき, 過去の履歴をどのように利用したら良いのか. 本発表ではこの問題を強化学習の枠組みで検討する.

非零和ゲームの例として日本の子供の遊びである１・２・５じゃんけんを取り上げる. １・２・５じゃんけんは, 二人で行うじゃんけんの一種で, グー (G), チョキ (C), パー (P) で勝つと, それぞれ 1,2,5 点を得るゲームである. このとき, 負けても点を失わないのが非零和版である. １・２・５じゃんけんの利得表を表 1 に示す.

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

表 1: １・２・５じゃんけんの利得表

このゲームでは, どのような戦略が考えられるだろうか. 明らかに特定の手を出す純戦略は最適戦略になり得ない. しかしながら混合戦略の範囲では, お互いに相手の手の最適戦略になっている Nash 均衡解が必ず存在する. このゲームでは Nash 均衡解は双方が G,C,P を確率 $(x,y,z)_N = (2/17, 10/17, 5/17)$ で出すことで, そのときの平均利得は 10/17 となる. また, 相手の手に関わらず自己の最低点を確保する手 (MinMax 戦略) としては, $(x,y,z)_M = (10/17, 5/17, 2/17)$が存在し, そのときの平均利得はやはり 10/17 である. 一方, 両者がランダムに手を選んだ場合 ($(x,y,z)_R = (1/3, 1/3, 1/3)$) 得られる平均利得は 8/9 となる. また双方が 1/2 の確率でランダムに G,P を出すと, 平均利得は 5/4 となる. このように, うまく協調できれば Nash 均衡解よりも双方が有利となる戦略があり得るのである.

[1]Reinforcement Learning of non zero-sun game — taking 1 2 5 janken as an example
Tsutomu Goto, Motoshi Iwata, Akira Ito, Kazunori Terada, Graduate School of Engineering, Gifu University

以上は一回毎の対戦を独立に考えたときの戦略であるが, このゲームを繰り返し行う場合は他の戦略も可能である. たとえばお互いに交替して P で勝つことにすれば, 5/2 の平均得点を得ることができる. しかしながら, 我々はエージェント同士が話し合って行動するのではなく, 独立して自己の得点を最大化するように行動するものと考える. 直接的な協調の手段が与えられていない状況で, どのようにして自己の最適行動を生成したら良いのか, これが本発表で検討する課題である.

2 履歴を用いた強化学習

ここでは強化学習の枠組みとしてQ学習を用いる. Q学習では, 過去に状況 S で自分が選択した行動 a とその時得られた報酬 r の組から, 現在の状況 S_t で行動 a_t をとるときの割引された報酬の和の期待値 $Q(S_t, a_t)$ を求める. しかしながら, 相手の戦略が決まらない (分からない) 限り自己の報酬の期待値は求まらない. そこで, 相手は過去 h 回の双方の手の組によって次の出す手を決めている「h 次マルコフ戦略」と仮定すると, 過去 h 回の相手と自分の手の組

$$\{a^m_\tau, a^o_\tau\}^{t-h}_{t-1} = \{a^m_{t-1}, a^o_{t-1}, a^m_{t-2}, a^o_{t-2}, \cdots, a^m_{t-h}, a^o_{t-h}\}$$

を状態と考えることで, 相手を含む世界を MDP(マルコフ決定過程) としてモデル化できる.

Q学習の標準的アルゴリズム [1] では, 状態 S_t で自分が a^m, 相手が a^o をとり, 報酬 $r(a^m, a^o)$ を得て, 状態 S_{t+1} に移行したとすると, $Q(S_t, a^m)$ の更新式は以下で与えられる.

$$Q(S_t, a^m) \leftarrow (1-\alpha)Q(S_t, a^m) + \alpha(r(a^m, a^o) + \gamma \max_{a^m} Q(S_{t+1}, a^m))$$

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行動 a^m を選択
・それ以外
$p(a^m) = C\exp(Q(S, a^m)/T)$ の確率で行動 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.

3 実験

1・2・5じゃんけんの振る舞いを調べるために, 履歴を用いたQ学習を行うエージェント同士を実際に対戦させてみる. 実験に用いた戦略は履歴長 h を用いるQ学習 SQh($h = 0, 1, 2, 3$) である. 比較のため G,C,P を等確率で選択する Random 戦略を含めて, リーグ戦を行った. 対戦結果を表 2 に示す.

学習に使用したパラメーターは, $\alpha = 0.1$, $T = 0.2$, $\gamma = 0.9$, $\epsilon = 0.01$, また Q の初期値は 0 とした. 各対戦は 10^8 回のじゃんけんを 1 試合とし, その最後の 10^7 回を収束値と考え平均値を求める. これを乱数を変えて 10 回行い平均を取る. 表の数値は, 左側の戦略が上側の戦略と対戦したときの左側の戦略の平均得点である.

	Random	SQ0	SQ1	SQ2	SQ3
Random	0.889	0.673	0.673	0.671	0.670
SQ0	1.631	0.879	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

表 2: Random, SQh のリーグ戦対戦結果

表から以下のことが分かる.

- 学習するエージェントは Random に対し, (常に P を出すことによって) 高い得点をあげている.
- 履歴を用いない SQ0 に対しては, 履歴を用いた戦略が高い得点をあげている.
- 履歴を用いた学習 (SQ1,SQ2,SQ3) 同士の対戦では, 混合戦略の範囲では実現できない高い得点をあげている.

SQ2-SQ2 対戦における平均得点の時間変化を図 1 に示す. 図から分かるように, 平均得点は 2.5 を上限に大きく振動する. これは, 確率 ϵ で行動がランダムに選択されるため, 一旦協調的行動が成立してもまた崩壊してしまうためである. 学習の進行に伴い ϵ を小さくすればこの現象は防げるが, 我々はむしろ協調の発現と崩壊の反復こそがエージェントにおける協調の本質であると考える. この機構があるからこそ, 搾取できる相手には搾取する, それができなければ協調を模索するなど, 相手に応じて柔軟な戦略を選択することができるのである.

図では, 5×10^7 回の時点で一方のエージェントの学習を停止させたものを併せて示してある. 一方 (P1) が学習を停止すると, 他方 (P2) はそれに対して最適化した戦略を獲得するため, 停止した方 (P1) の平均利得は

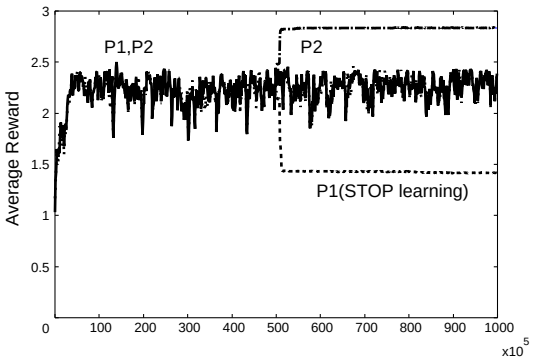

図 1: SQ2-SQ2 対戦での平均得点の時間変化

低く, また学習を続けている方 (P2) の利得は高くなる. 学習を止めることは学習エージェント同士の対戦では致命的なのである.

協調成立における ϵ の影響を調べるため, 履歴長 0,1,2,3 同士の対戦で ϵ を変えたときの平均得点を図 2 に示す. 図から分かるように, 適切な ϵ を選ぶことで, 平均利得を 2.5 に近づけることができる. 一般に ϵ が小さいほど, 一旦成立した協調を破壊することが少ないため, 平均得点が高くなる. ただし ϵ をあまり小さくすると探索の速度が遅くなる. 実際, 探索空間の大きな $h = 3$ では 10^8 回では探索が完了せず, 小さな ϵ での性能は良くない.

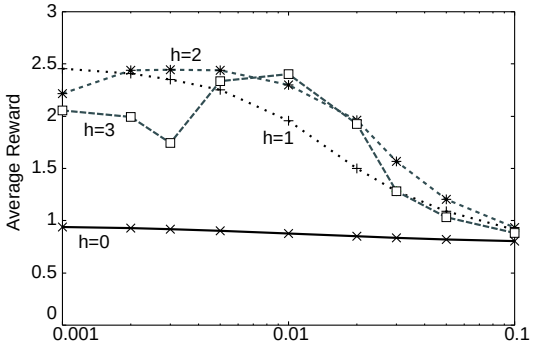

図 2: ϵ を変化させた時の平均得点の変化

4 まとめ

我々は, 適切な長さの履歴を用いるQ学習により, 繰り返し非零和ゲームに強化学習が適用できることを示した. また, エージェントは自己の利益を追求する過程で, 協調的行動を発現させ得ることを示した. この時, 学習を続けることが協調の維持にとって本質的であることを示した.

参考文献

[1] Sutton, R.S. and Barto, A.G.: Reinforcement Learning, The MIT Press, 1998.