6R-5

非ゼロ和ゲームにおけるエントロピーを基にした相手戦略の推定

大橋 資紀† 伊藤 昭† 寺田 和憲†
岐阜大学大学院工学研究科†

1. はじめに

利害の対立する状況でのインタラクションでは，適切な相手モデルを構築し，それを用いた戦略をとる必要がある．相手が固定戦略ならば，相手の行動の統計モデルを相手モデルとすればよい．しかし，相手が固定戦略ではなく，行動を変化させることのできる学習戦略の場合，統計モデルでは不十分である．たとえば相手の戦略を，これまでの行動履歴を基に自分の行動を決定し，過去と同じ状況でも，異なった行動を生成し得る学習戦略だとしよう．このような場合，統計モデルは明らかに不適切である．なぜなら，観測した相手モデルは過去のモデルであって，現在の相手モデルではなく，ましてや未来を予測するモデルにはなり得ないからである．

特に，繰り返し非ゼロ和ゲームでは，互いに敵対するよりも協調行動をとることによって，多くの利益を得られる場合がある．このような状況では，相手が協調的なのか敵対的なのかを判断する必要があるが，自分が相手に敵対的／協調的であることが相手を敵対的／協調的にするかも知れず，相手のモデル化には慎重を要する．

以上のようなことから，統計モデルではない相手モデルの構築が必要と考え,「学習戦略」という制約の下での相手モデルの推定を行い，行動決定に利用することを考える．

我々は手始めに，学習戦略で重要な役割を果たすものとして，相手が行動決定に用いている行動履歴のサイズ（履歴長）を推測することを試みる．また，具体的に「125 じゃんけん」の例を用いて，推定した相手モデルを用いた行動選択の優位性を示す．

Entropy-Based Strategy Estimation in The Non-zero-sum Game

†Motoki Ohashi, Akira Ito, Kazunori Terada, Gifu Univ.

2. エントロピーを用いた相手履歴長の推定

両プレイヤは，何らかの方針に従って自己の行動を選択する必要があるが，ここでは「過去 k ステップまでの，自分と相手の行動履歴（履歴長 k の行動履歴）を基に，自己の行動を決定する」と仮定する．例えば，過去の履歴に無関係に行動を選択しているのであれば履歴長 0，過去 1 ステップ前と 2 ステップ前の履歴を基にしているのであれば履歴長 2 となる．

ここで以下で使用する記号の形式的定義を与える．履歴長 k の履歴とは，ステップ t のときの自分の行動を a_t^m，相手の行動を a_t^o として，$S_t^{(k)} = \{a_{t-1}^m, a_{t-1}^s, \ldots, a_{t-k}^m, a_{t-k}^s\}$ である．

エージェントは，各履歴長 $k(k = 1, 2, ..., k_{max})$ について，履歴 $S^{(k)}$ に対する相手の行動確率 $p(a^o|S^{(k)})$ を，次式に従って求めておく．

$$p(a_t^m, a_t^o|S_t^{(k)}) \leftarrow p(a_t^m, a_t^o|S_t^{(k)}) + \delta$$

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

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

次に，履歴 $S_t^{(k)}$ に対する相手の行動の不確実度を表すエントロピー $H(S_t^{(k)})$ を確率 $p(a^o|S_t^{(k)})$ を用いて次式により求める．

$$H(S_t^{(k)}) = -\sum_{a^o} p(a^o|S_t^{(k)}) \log p(a^o|S_t^{(k)})$$

これは，エントロピーの経路平均に相当する．この値は振動が激しいので以下の式で $H(S_t^{(k)})$ 平均して，最終的なエントロピー $H(k)$ を求める．

$$H(k) = (1-\delta)H(k) + \delta H(S_t^{(k)})$$

協調可能なときは履歴長 1 を用いて協調を、そうでなければランダムな手を出す固定戦略（N 戦

略）と対戦したときのエントロピー $H(k)$ を，図 1 に示す。協調未成立のときにはどの k についても高い値であるが、協調が始まると $k \geq 1$ のエントロピーは急速に減少し、相手が履歴長 1 で行動し始めたことを示す。

一般に相手の履歴長以上のエントロピーは小さく，それ未満が大きくなる．このことを利用し，有効履歴長 k_E を次式で求める．

$$k_E = \frac{\sum_{k=0}^{k_{max}} (H(k) - H(k_{max}))}{H(0) - H(k_{max})}$$

ここで求めた有効履歴長 k_E が相手の履歴長の推測値となる．

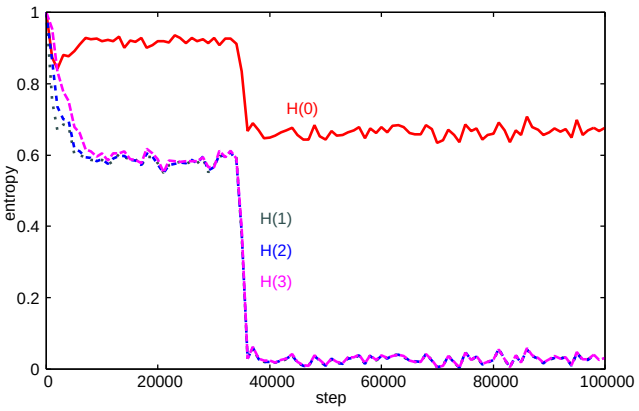

図 1: N 戦略と対戦したときのエントロピー

3. 提案手法の検証

Q 学習を拡張させた確率学習（PQ）を行うプレイヤと，PQ に有効履歴長の推定機能を付加し，有効履歴長の履歴を用いて行動を決定するプレイヤを対戦させ，提案手法の検証を行った．検証には表 1 のような利得の 125 じゃんけんを用いた．

表 1: 125 じゃんけんの利得

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

このゲームでは，ランダムに手を選択した場合，平均 8/9 点である．相互に相手の手の最適戦略となっている Nash 均衡解は，G:C:P を 2/17:10/17:5/17 の確率で選択する混合戦略であり，平均得点は 10/17 点である．これらに対して，お互いに G と P を交互に出す戦略（協調戦略）をとれば，双方とも平均 2.5 点を得る．

この 125 じゃんけんを，$k_{max} = 3$ で相手履歴長の推定を行う PQ と履歴長 2 の通常の PQ が対戦した場合，図 2 のように相手履歴長の推定を行った PQ の方が得点が高い．同時に，協調戦略も実現できている．

このときの各履歴長のエントロピーと有効履歴長を図 3 に示す．有効履歴長は 1〜2 のあたりにあり，うまく推定できていると言える．従って，推定した有効履歴長を用いた方が，用いない場合よりも優位であると言える．

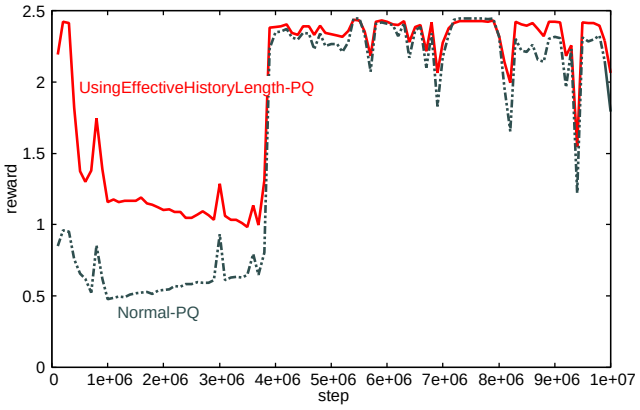

図 2: 得点推移

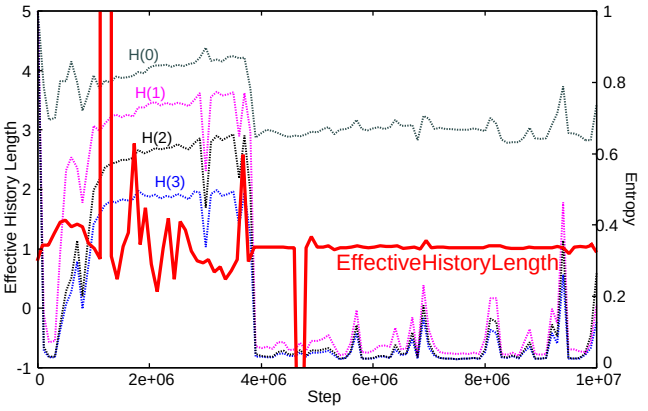

図 3: エントロピーと有効履歴長の推移

4. まとめ

繰り返し二人非ゼロ和ゲームにおいて，エントロピーを用いて相手の履歴長を推定する手法を提案し，推定した履歴長を利用することの優位性を示した．本論文では，相手履歴長の推定に留まったが，相手が学習戦略であるか固定戦略であるかの推定や，相手の学習速度の推定を加えた相手モデルを用いることで，長期的な視点での相手予測が行うことができ，様々な相手に対して有利な戦略をとることが可能になると思われる．