空間上の囚人のジレンマゲームにおけるエージェントの進化・学習

小野真裕

1　はじめに

近年，物理的実体の有無に関わらず自律したエージェントの研究が盛んである．エージェント同士がコミュニケーションする場合，同一の目的に対し協働する場合もあれば，利害関係が相反し競争状態に陥る場合もある．どちらの状況も人間社会においても古くから存在してきたが，人間の持つウェットな感情を排したコンピュータ上のコードでは，より明確な形で表れる．特にネット上で分散したエージェントに対しては一般に拘束力が存在しないため，独立したエージェント同士は非協力関係になりやすい．ここで，エージェント同士のジレンマ状況を乗り越えて協働状況を自然に作る仕組みは，マルチエージェントシステム（全てのエージェントを含む系）のパフォーマンスを向上させる上で非常に重要となってくる．

エージェントは，誕生と同時に既にプログラムされた基本的な行動選択機能を持つのは当然のことながら，さらに学習部分をもつことで環境にうまく適応することができるようになる．そのため，システムのパフォーマンスは，ハードコーディング部分と，環境に適応可能な学習部分との組み合わせ方に影響を受ける事は容易に予想される．マルチエージェント環境における学習の効果を明らかにすることはエージェント設計指針のために有益である．

システムを調べるにあたっては，システムの構成要素間の相互接続構造，構成要素の動作，構成要素間の動作に対する相互反応が重要である．本研究では，エージェント間の相互作用が空間構造によって制限される場合に特に注目し，エージェントの進化・学習について報告する．要素のネットワーク構造は small-world ネットワークモデル [1] を，要素間の動作に対する相互作用は非協力ゲームの囚人のジレンマゲームをとりあげている．

2　エージェントモデル

エージェントモデルを説明する．ここでいうエージェントは繰り返し囚人のジレンマゲームをプレイするプレーヤーであり，以降プレーヤーと呼ぶ．プレーヤーは他のプレーヤーとの対戦の過程で学習するための情報を遺伝子として持ち，強化学習を行う．

繰り返しゲームを行う際，プレーヤーは過去一回のゲームの記憶を保持し，前回の自分と相手の手番の組み合わせによって状態を認識する．プレーヤーは全ての状態・手番の組み合わせについて価値関数を保持し，ある状態 s において行う手番 a を価値関数 $Q(s,a)$ によって評価する．$Q(s,a)$ が大きいほど a は優れていると評価された手番であり，プレーヤーは優れた手番を選択するべきである．ただし，一般にプレーヤーは行動探索的な機構も備えているべきであり，本モデルでは次の手番 a_{t+1} は ϵ グリーディによって選択する．

選択した手番に対しては，得られた利得を報酬として価値関数が更新される（式 (1)）．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \quad (1)$$

ただし，α は学習率，r_{t+1} は手番 a_t に対する報酬，γ は割引率である．学習モデルとしては本研究では sarsa[2] を採用した．遺伝子には，これら強化学習のためのパラメータと価値関数の初期値が含まれる．

3　シミュレーション実験

実験では，n 人のゲームプレーヤー集団を作成した後，以下の各世代の処理を g 回行い集団を進化させる．

i) 空間構造に依存して選択されたプレーヤー同士で，a 回繰り返しゲームを単位とする対戦を行う．ii)

その世代における全ての対戦の終了後，各プレーヤーがその世代において取得した利得に応じて，b%のプレーヤーを確率的に死亡させる．iii) 空間構造に依存した方法で集団が n 人になるようにプレーヤーを補充する

ここで，空間に構造が存在しないプールケースと，プレーヤー間にネットワークが存在するネットワークケースを考える．それぞれに依存する処理は以下である．プールケースでは，i) 各プレーヤーについてランダムに対戦相手を探し m 回の対戦を行う．iii) 2 人トーナメント，2 点交差，c%突然変異を行い母集団を再構成する．ネットワークケースでは Small World ネットワークモデル [1] を用い，ランダムネスをシミュレーションパラメータとする．各世代の処理は以下である．i) 各プレーヤーが持つ m 本のリンクで結ばれるプレーヤー同士で対戦を行う．iii) 死亡したプレーヤーの位置からリンクで接続された近傍のプレーヤーのうち最も利得の大きいものを c%突然変異を考慮しコピーする．

本稿では，$(a, b, c, m, n) = (100, 20, 0.2, 3, 400)$ とし，プレーヤーパラメータの一つ $\epsilon = 0.1$ と固定した．囚人のジレンマの利得行列は，$(T, R, P, S) = (5, 3, 1, 0)$ とした．また，ある確率でプレーヤーの選択した手番が逆転するノイズという概念を取り入れている．

3.1 シミュレーション結果

図 1 にプールケースの，図 2 にネットワークケースのプレーヤーパラメータ α，γ を示す．ノイズに対する変化が顕著でないため，ケースの違いにおける差異を見るべく平均して特徴量とすると，プールケースにおいては $\alpha \simeq 0.62$，$\gamma \simeq 0.19$，ネットワークケースでは $\alpha \simeq 0.36$，$\gamma \simeq 0.50$ であった．ここで，α はプレーヤーの価値関数の更新速度を，γ はプレーヤーが未来を重視する程度を意味する．他プレーヤーと再度対戦することが多いネットワークケースでは，プールケースに比べ割引率が高く，未来に得られる利得を重視している．また，価値関数の更新速度も比較的緩やかである．それに対して，プールケースのように他プレーヤーとの関係の継続性が希薄な場合には，直近の報酬しかあてにせず価値関数の更新速度が速い．このように空間構造によって学習のプロセスに違いがあることがわかった．

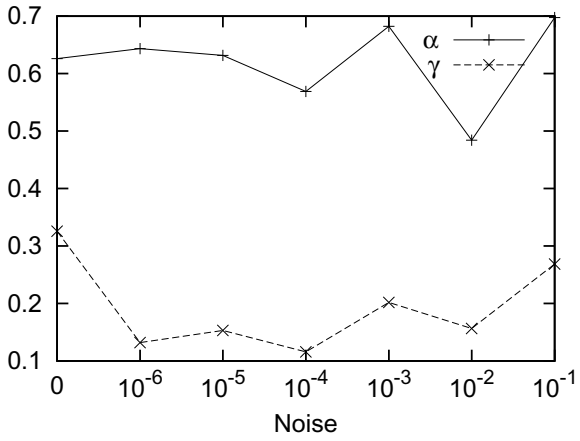

図 1: α，γ（プールケース）

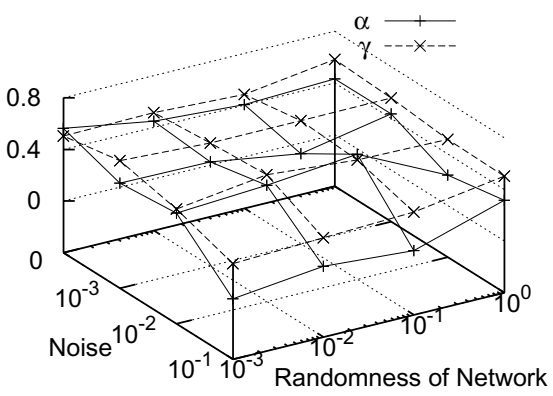

図 2: α，γ（ネットワークケース）

4 おわりに

空間構造がエージェントの進化・学習に与える影響について調べた．その結果，プレーヤー間の関係の継続性に依存して学習率・割引率に変化が現れることがわかった．

参考文献

[1] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, Vol. 393, pp. 440–442, June 1998.

[2] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning.* MIT Press, 1998.