1C-06

クールノー競争におけるマルチエージェント強化学習に関する研究

豊島 健太郎 * 坂本 充生 † 阿部 拳之 ‡ 岩崎 敦 †
Kentaro Toyoshima Mitsuki Sakamoto Kenshi Abe Atsushi Iwasaki

1 はじめに

本研究ではクールノーゲームにおいて，個々のエージェントが強化学習アルゴリズムに従った場合，どのような振る舞いを学習するか分析する．人がどう協力するかの仕組みは人工知能，経済学，生物学等における学際的な研究課題である．不完全観測下の無限回繰り返しゲームおいてどのような戦略が均衡になるかは十分にわかっていない．戦略空間を有限状態機械に限定すると，見間違えが起きても協力状態を回復しやすいシンプルな戦略が生き残ることが知られているが，複雑な戦略を含めた分析はまだまだ困難である [4].

近年，企業は戦略的な決定をアルゴリズムに委ねるようになってきている [1]．その中には，機械学習のような能動的に学習するものも含まれ，競合する企業同士がそのようなアルゴリズムで意思決定を行うと，人間の介入なしに談合を学習することがある．例えば，Amazon 等の大手 EC サイトでは，個人や他社の価格に応じて商品の値段をつけることで，利益を増加する仕組みがあり，電力市場では，他社の価格設定がわからない状態でより多くの利益をするアルゴリズムが導入されている．これらの市場には，クールノー競争と呼ばれる複数の企業が支配する寡占市場における財の供給量から市場価格と利益が与えられる経済学のモデルが適している．本研究では，市場を支配する企業が 2 社のみの複占市場における同質財の供給量を決定するようなクールノー競争のモデルを考える．計算実験の結果，エージェントは非協力的な戦略を学習することがわかった．

2 クールノーゲーム

複占市場におけるプレイヤ $i \in \{1,2\}$ が供給量を決定するモデルを考える．本章では文献 [2] にもとづいて，2 人不完全観測下の無限回繰り返しゲームをモデル化する．ここでプレイヤ i は成分ゲームを無限期間 $t = 0,1,2,\cdots$ に渡って繰り返す．プレイヤ i は毎期 t に供給量 $q_{it} \in Q$ を決める．プレイヤの供給量と確率的に決まる需要量 $d_t \in D$ から価格 $p_t = d_t - (q_{1t} + q_{2t})$ が決定する．さらに，価格 p_t と供給量 q_{it} の積で，プレイヤ i の利益 g_{it} が決定する．

無限回繰り返しゲームでは割引因子 δ が定める確率で次もゲームが継続するかを定める．不完全観測下のクールノー競争では，相手の供給量 q_{-it} は観測できず，自分の供給量 q_{it} と市場価格 p_t のみを観測する．

表 1: クールノー競争の利得表

	$q_2 = q^M$	$q_2 = q^C$
$q_1 = q^M$	18, 18	15, 20
$q_1 = q^C$	20, 15	16, 16

次に，談合供給量と均衡供給量について説明する．d_t は確率的であるため，その期待値 $\overline{d_t}$ を用いる．談合供給量は，プレイヤ間の談合によってお互いの利益を最大にするような供給量であり，$q_{1t} = q_{2t} = \overline{d_t}/4$ で与えられる．均衡供給量は，お互いに談合から逸脱した場合であり，供給量は増え $q_{1t} = q_{2t} = \overline{d_t}/3$ となる．クールノーゲームでは，均衡供給量を互いに選択すると談合するよりも利益が小さくなるが，談合から逸脱することが均衡の選択になる．例えば，表 1 のように $\overline{d_t} = 12$ のとき，談合供給量と均衡供給量はそれぞれ $q^M = 3, q^C = 4$ となる．g_{it} = お互いが q^M を選択すれば，高い利益を獲得できる．しかし，一方が談合から逸脱することで，そのプレイヤは協力していたときよりも高い利益を獲得できるが，他方は利益が小さくなる．そのため，互いに q^C を選択することが均衡となる．

クールノー競争は囚人のジレンマの一般的なモデルであり，1 回限りではお互いに協力（談合）することで，高い利得を獲得できるが，互いに裏切り合う（逸脱）ことが均衡となる．しかし，長期間であれば協力関係を築くことがある．

本研究では，文献 [1] に基づいて，需要量を設定する．需要量 d_t は $d^1 = 287.5, d^2 = 312.5$ の 2 種類とする．この場合，$\overline{d} = 300$ より，談合供給量 $q^M = 75$，均衡供給量 $q^C = 100$ となる．本研究では，クールノー競争でもこの繰り返し囚人のジレンマのように談合を実現するかを吟味する．

3 学習アルゴリズム

強化学習は，エージェントが試行錯誤により適切な振る舞い (方策) を学習する．エージェントはある時刻 t において状態 $s_t \in \mathcal{S}$ を観測する．次に行動 $a_t \in \mathcal{A}$ を方策 $\boldsymbol{\pi} : s \to \Delta\mathcal{A}$ に従い決定する．行動に応じて，報酬 $r_t \in \mathbb{R}$ と新たな状態 s_{t+1} を受け取る．この報酬 r_t から方策 $\boldsymbol{\pi}$ を更新し，適切な振る舞いを学習する．マルチエージェント強化学習は，複数のエージェントが同時に方策を学習するため，シングルエージェン

* 電気通信大学情報理工学域
† 電気通信大学大学院情報理工学研究科
‡ 株式会社サイバーエージェント

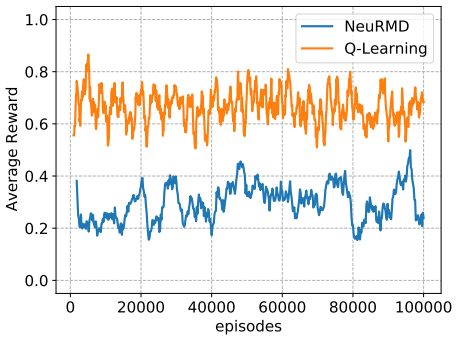

図 1: 獲得報酬の時間平均の推移

トの学習と比較して様々な問題が存在する．マルチエージェント強化学習では，時刻 t において複数のエージェントがそれぞれ状態 $s_{t,i}$ を観測し，行動 $a_{t,i}$ を決定する．そのため，エージェント i が受け取る報酬 $r_{t,i}$ と次の状態 $s_{t+1,i}$ は，自分の行動 $a_{t,i}$ だけでなくその他のエージェントの行動 $a_{t,-i}$ の行動によって決定される．また他のエージェントも同時に方策を学習するため，振舞いが非定常的に変化する．

プレイヤが観測する状態は自分の供給量と市場価格の過去 1 期の履歴に限定し，$\mathcal{S} = \{(q_{i,t-1}, p_{t-1})_{\mathcal{Q},\mathcal{P}}\}$ とする．プレイヤが選択できる供給量は連続値のため，談合供給量 $q^M = d_t/4$ と均衡供給量 $q^C = d_t/3$ の間を離散化することで，行動集合を定義する:

$$\mathcal{Q} = \{q^1, q^2, ..., q^k\} \text{ s.t. } q^{j+1} - q^j = v.$$

需要量 d_t は h 個の値の集合 $\{d^1, d^2, ..., d^h\}$ s.t. $d^{j+1} - d^j = mv$ の中からランダムに選ばれる．獲得する報酬 r_t は $r_t := g_{it}$ を標準化した値である．

先行研究 [1] で用いられていた手法は，Q-Learning を用いた ϵ-greedy 法と呼ばれる手法である．ϵ-greedy 法は確率 $\epsilon \in [0, 1]$ でランダムに行動を選択し，それ以外は効用関数が最大になる行動を選択する確率的方策である [3]．探索確率 ϵ は，時間ステップ数に応じて徐々に値を小さくして，探索の頻度を減らし，データ活用を行うようにすることで，探索と活用のバランスを調整する．

次に本研究で用いた Actor-Critic 手法の 1 つである NeuRMD を概説する．Actor-Critic 法ではロジット $\boldsymbol{y}$ を方策のパラメータ $\boldsymbol{\theta}$ によって，$\boldsymbol{y} = (\theta(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ のように表現し，ロジット $\boldsymbol{y}$ からソフトマックス関数で各状態でどの行動をとるかを規定する:

$$\pi(a|s) = \frac{\exp(y(s,a))}{\sum_{a'\in\mathcal{A}} \exp(y(s,a'))}.$$

本研究で用いた Actor-Critic 手法である NeuRMD ではロジット $\boldsymbol{y}$ を

$$y_{t+1}(s,a) = y_t(s,a) + \eta\left\{A^\pi(s,a) + \frac{u}{\pi(a|s)}\left(\frac{1}{|\mathcal{A}|} - \pi(a|s)\right)\right\}$$

にしたがって更新する．ここで，η は学習率，u は突然変異率とする．一般的な方策勾配法 [3] ではロジット $\boldsymbol{y}$ を

$$y_{t+1}(s,a) = y_t(s,a) + d^\pi(s)\eta\pi(a|s)A^\pi(s,a) \quad (1)$$

にしたがって更新する．ここで，$d^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^\infty \delta^t P(s_t = s|s_0, \boldsymbol{\pi})]$ は期待割引累積訪問数とする．

4 計算機実験

行動数 15 のクールノーゲームにおいて，NeuRMD の獲得報酬を確認し，Q-Learning と比較する．行動集合は $\mathcal{Q} = \{70, 72.5, ..., 102.5, 105\}$ s.t. $v = 2.5, m = 8$，割引因子 $\delta = 0.95$，学習率 $\eta = 0.02$，NeuRMD の突然変異率は $u = 0.01$ とする．δ によるゲームの終了までを 1 エピソードとして，100,000 エピソード学習し，ランダムシードを変えながら行った 3 試行を評価する．

図 1 は各手法の平均報酬値の推移を示す．平均報酬値は，そのエピソードの 2 体の 1 ゲームあたりの平均報酬値とした．NeuRMD における獲得報酬は 0.30 付近で推移していたが，Q-Learning は 0.65 付近で推移した．このことから，NeuRMD の挙動は，時間減衰する Q-Learning に比べ，非協力的な方策を学習することがわかった．

次に，それぞれの手法が学習した方策に対して近似的に最適反応方策を求めることで，均衡に近い方策か評価した．具体的には，学習した方策の組 (π_0, π_1) の片方の方策 π_i を固定し，クールノー競争を方策勾配法で学習させることで，近似的に最適反応方策 π_{-i}^{BR} を得る．方策 π_i に対して，学習した方策 π_{-i} が獲得する報酬と，近似的な最適反応方策 π_{-i}^{BR} の獲得する報酬を比較する．この報酬差の合計が小さいほど，π_{-i} が均衡に近いといえる． この結果，Q-Learning は報酬差の平均値が 0.274 だったのに対し，NeuRMD では 0.0355 と，より小さい値となった．すなわち，方策勾配法が学習した方策と比べると NeuRMD のほうがより高い報酬値を獲得することを確認した．このことから，NeuRMD が Q-Learning より均衡に近い戦略を学習することを示せた．

参考文献

[1] E. Calvano, G. Calzolari, V. Denicoló, and S. Pastorello. Algorithmic collusion with imperfect monitoring. *International Journal of Industrial Organization*, 2021. To appear.

[2] ヨンジュン, 岩崎, 神取, 小原, 横尾. 部分観測可能マルコフ決定過程を用いた私的観測付き繰り返しゲームにおける均衡分析プログラム. 情報処理学会論文誌, pp. 1234–1246, 2012.

[3] 森村. 強化学習. 講談社, 2019.

[4] 西野上, 五十嵐, 岩崎. 私的観測下の繰り返し囚人のジレンマにおける協力のダイナミクス. 第 19 回情報科学技術フォーラム, 2020.