日本オペレーションズ・リサーチ学会
2021年 秋季研究発表会

見間違えのある繰り返しゲームのための Actor-Critic 型強化学習

電気通信大学 *坂本 充生 SAKAMOTO Mitsuki
株式会社サイバーエージェント 阿部 拳之 ABE Kenshi
05000493 電気通信大学 岩崎 敦 IWASAKI Atsushi

1. はじめに

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．見間違えのある（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振る舞い (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 [5]．そこで，個々のエージェントが強化学習にしたがうダイナミクスが，どんな振る舞いを学習するか吟味する．

本研究では，見間違えのある繰り返し囚人のジレンマというマルチエージェント系の上に，学習アルゴリズムを表現する枠組みを定義する．既存手法として，Actor-Critic 型強化学習の 1 つである q-based Policy Gradient (QPG) [2] と Neural Replicator Dynamics (NeuRD) [1] を扱う．これらに対して突然変異付きレプリケータダイナミクスに着想したアルゴリズムである Neural Replicator-Mutator Dynamics (NeuRMD) を提案する．3 つの手法それぞれでエージェントがどんな振る舞いを学習するか吟味した結果，NeuRMD がもっとも高い利得を実現する方策を学習しやすいことがわかった．さらにその方策は Win-Stay, Lose-Shift (WSLS, 図 2) という均衡戦略に相当することもわかった．

2. モデル

本章では文献 [4] にもとづいて, 2 人私的観測付き無限回繰り返しゲームをモデル化する．ここでプレイヤ $i \in \{1,2\}$ は成分ゲームを無限期間 $t = 0, 1, 2, \cdots$ に渡って繰り返す. 各期においてプレイヤ i は有限集合 A から行動 a_i を選択し, その行動の組を $\boldsymbol{a} = (a_1, a_2) \in A^2$ とする. 次に, プレイヤ i は $\boldsymbol{a}$ に関する私的なシグナル $\omega_i \in \Omega$ を観測する. $\boldsymbol{w}$ をシグナルの組 $(\omega_1, \omega_2) \in \Omega^2$ とする. また, プレイヤが $\boldsymbol{a}$ を選択したとき $\boldsymbol{w}$ が生起する同時確率を $o(\boldsymbol{w} \mid \boldsymbol{a})$ とし, この同時確率を与える分布のことをシグナル分布と呼ぶ. プレイヤ i の利得は $g_i(\cdot)$ の値は表 1 によって定められた値に従う.

このとき，有限集合 Ω に対する $o_i(\omega_i \mid \boldsymbol{a})$ を Ω の限界分布 (marginal distribution) とする．加えて，どのプレイヤも他のプレイヤが選択した (または選択しなかった) 行動を正確には分からないと仮定する．これに合わせて各期の “実現利得 (realized payoff)” をプレイヤ i の行動 a_i とシグナル ω_i から決定する．具体的には $g_i(\boldsymbol{a}) = \sum_{\boldsymbol{w} \in \Omega^2} r_i(a_i, \omega_i) o(\boldsymbol{w} \mid \boldsymbol{a})$ を満たす $r_i(a_i, \omega_i)$ とする．

表 1: 囚人のジレンマ ($g > 0$, $l > 0$, および $|g - l| < 1$)

	$a_2 = C$	$a_2 = D$
$a_1 = C$	$1, 1$	$-l, 1 + g$
$a_1 = D$	$1 + g, -l$	$0, 0$

次にシグナル分布 $o(\boldsymbol{w} \mid \boldsymbol{a})$ で規定する，プレイヤ 2 の行動に関するプレイヤ 1 のノイズを含む観測をプレイヤ 1 の私的シグナルとし，$\omega \in \{g, b\}$ (*good*, *bad*) とする．正しい観測ではプレイヤ 2 が C を選択した際のプレイヤ 1 の私的シグナルは g, D を選択した際は b となる. これはプレイヤ 2 についても同様である.

3. マルチエージェント強化学習

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.

エージェントが観測する状態は，過去 1 期の履歴に限定し，$\mathcal{S} = \{\emptyset, Cg, Cb, Dg, Db\}$ とする．ここで $\emptyset$ はゲームの始まりにおける状態を指す．また報酬から相手の行動を推定させないため，報酬 r には実現利得 $r(a, \omega)$ を与える．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.

表形式で方策関数を表現する Actor-Critic 法では，Critic が各 (s, a) に対するアドバンテージ

$$A^{\boldsymbol{\pi}}(s, a) = Q^{\boldsymbol{\pi}}(s, a) - \textstyle\sum_{a'} \pi(a'|s) Q^{\boldsymbol{\pi}}(s, a')$$

を用いて現在の方策を評価し，そこから Actor が方策を更新する．ただし，$Q^{\boldsymbol{\pi}}(s, a) = E_{\boldsymbol{\pi}}\left[\sum_{k=t}^{\infty} \delta^{k-t} r_k | s_t = s\right]$ とする．また表形式ではロジット $\boldsymbol{y}$ を方策のパラメータ $\boldsymbol{\theta}$ によって,

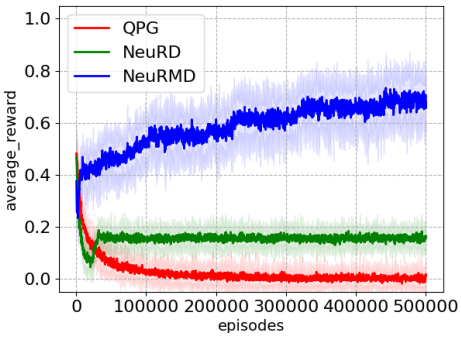

図 1: 2 体の平均獲得報酬の推移

$\boldsymbol{y} = (\theta(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ 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:

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

それぞれの手法でロジット $\boldsymbol{y}$ をどのように更新するかを概説する．まず，QPG では

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

にしたがって更新する．ここで，η は学習率，$d^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{\pi}}[\sum_{t=0}^{\infty}\delta^t P(s_t = s|s_0,\boldsymbol{\pi})]$ は期待割引累積訪問数とする．次に NeuRD [1] では

$$y_{t+1}(s,a) = y_t(s,a) + d^{\boldsymbol{\pi}}(s)\eta A^{\boldsymbol{\pi}}(s,a) \qquad (2)$$

にしたがって更新する．これは QPG とは独立にレプリケータダイナミクスからロジットの更新式 1 を導いている．最後に提案手法である NeuRMD では

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

にしたがって更新する．ここで，u は突然変異率とする．これは突然変異付きレプリケータダイナミクス [3] から，NeuRD と同じようにロジットの更新式 2 を導いている．また，突然変異項を追加するだけでなく，期待割引累積訪問数を取り除くことで相互協力状態を学習しやすいようにしている．

4. 計算機実験

本節では，囚人のジレンマの利得パラメータを $g = 0.1, l = 0.1$，割引因子 $\delta = 0.9$，シグナル分布のパラメータは $p = 0.95, q = 0.01$，アドバンテージの学習率 $\alpha = 0.1$, 方策の学習率 $\eta = 0.02$, 割引率 $\gamma = 0.9$, NeuRMD の突然変異確率は $\mu = 0.01$ とする．δ による

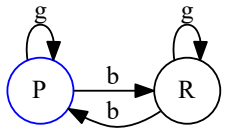

図 2: 状態 P からスタートする WSLS

ゲームの終了までを 1 エピソードとして，500,000 エピソード学習し，ランダムシードを変えながら行った 30 試行を評価する．

図 1 に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．縦軸は平均報酬値，横軸はエピソード数に対応し，区間数 10 の移動平均と信頼区間をプロットした．最終的な平均獲得報酬は，QPG が 0.014, NeuRD が 0.16，NeuRMD が 0.67 となり，NeuRMD が従来手法よりかなり高い報酬を与える方策を学習した．

この実験で NeuRMD が獲得した方策の多くは $\pi(C|\emptyset)$，$\pi(C|Cb)$ と $\pi(C|Dg)$ が 0.01，$\pi(C|Cg)$ と $\pi(C|Db)$ が 0.99 でとなっていた．興味深いことにこれは状態 P からスタートする WSLS とほぼ一致する (図 2)．WSLS は見間違えのある環境で最も高い利得を実現する均衡戦略として知られている [5]．このことから NeuRMD が非自明な協力的均衡戦略を従来手法より安定的に学習することを実験的に示せた．

参考文献

[1] D. Hennes, D. Morrill, S. Omidshafiei, R. Munos, J. Perolat, et al. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *AAMAS*, pp. 492–501, 2020.

[2] S. Srinivasan, M. Lanctot, V. Zambaldi, J. Pérolat, K. Tuyls, R. Munos, and M. Bowling. Actor-critic policy optimization in partially observable multiagent environments. In *NIPS*, pp. 3426–3439, 2018.

[3] B. Zagorsky, J. Reiter, K. Chatterjee, and M. Nowak. Forgiver triumphs in alternating prisoner's dilemma. *PLOS ONE*, pp. 1–8, 2013.

[4] 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.

[5] 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.