5C-05

二人零和ゲームにおける突然変異付きレプリケータダイナミクスを用いた学習アルゴリズムに関する研究

坂本 充生 * 阿部 拳之 † 岩崎 敦 *
Mitsuki Sakamoto Kenshi Abe Atsushi Iwasaki

1 はじめに

本研究では，二人零和ゲームにおける突然変異付きレプリケータダイナミクスを利用したオンライン学習アルゴリズムの帰結を吟味する．人工知能の分野における零和ゲームのナッシュ均衡学習は，将棋や囲碁をはじめ，ミニマックス問題を含む様々な分野での応用が期待される [1]．オンライン学習では，プレイヤは時刻ごとに行動の選択と利得 (損失) の観測を行い，それを基に戦略の更新を行う．損失関数の観測方法には，すべての行動の関数値を観測する完全情報フィードバック設定と，1つの行動の関数値のみを観測する部分的フィードバック設定がある．2 人のプレイヤは，お互いに行動を強制せず，コミュニケーションを取ることができないため，均衡戦略の学習は困難である．既存アルゴリズムの多くは完全情報フィードバック設定で，戦略の時間平均をとって初めて均衡に収束する．この収束保証は，時間平均の計算が困難なニューラルネットワークを用いた手法 (敵対的生成ネットワークなど) への応用に適していない．また，より現実に則した部分的フィードバック設定では，更新する戦略自体がナッシュ均衡に収束するような更新則を与えることはより困難となる．

二人標準形零和ゲームにおけるオンライン学習アルゴリズムを実験的に評価する．従来手法として時間平均戦略がナッシュ均衡へ収束することが保証される Follow the Regularized Leader (FTRL) [5] と，完全情報フィードバック設定で更新する戦略自体がナッシュ均衡へ収束することが保証される Optimistic Follow the Regularized Leader (OFTRL) [4] を扱う．これに対して突然変異に着想したアルゴリズムである Mutant Follow the Regularized Leader (MFTRL) を提案する．完全情報フィードバック設定と部分的フィードバック設定で 3 つの手法がどのような振る舞いを学習するか吟味した．実験の結果，2 つの設定において MFTRL のダイナミクスが時間平均を取らずに均衡に収束することを示した．

2 二人零和ゲーム

二人零和ゲームでは，プレイヤ $i \in \{1,2\}$ (相手プレイヤは $-i$) は有限行動集合 A から行動 a_i を戦略 $\pi_i \in \Delta A_i$ に従い決定する．その行動の組を $\boldsymbol{a} = (a_1, a_2) \in A = (A_1 \times A_2)$，戦略の組を $\boldsymbol{\pi} = (\pi_1, \pi_2)$ とする．プレイヤ i の利得は利得関数 $r_i : A \to \mathbb{R}$ で与えられ，$r_1(\boldsymbol{a}) = -r_2(\boldsymbol{a})$ を満たす．

戦略の組 $\boldsymbol{\pi}$ がナッシュ均衡にどれだけ近いかを測る指標に exploitability がある [3]．プレイヤが戦略の組 $\boldsymbol{\pi}$ に従うとき，プレイヤ i の行動 a_i に対する期待利得を Q 関数を用いて $Q_i^{\boldsymbol{\pi}}(a_i) = \mathbb{E}_{a_{-i}\sim\pi_{-i}}[r_i(a_i, a_{-i})] = \sum_{a_{-i}\in A_{-i}} r_i(a_i, a_{-i})\pi_{-i}(a_{-i})$ とする．またプレイヤ i の期待利得を価値関数を用いて，$V_i^{\boldsymbol{\pi}} = \mathbb{E}_{\boldsymbol{a}\sim\pi[r_i(\boldsymbol{a})]} = \mathbb{E}_{a_i\sim\pi_i}[Q_i^{\boldsymbol{\pi}}(a_i)] = \sum_{a_i\in A_i} Q_i^{\boldsymbol{\pi}}(a_i)\pi_i(a_i)$ とする．二人零和ゲームにおける exploitability は次のように定義される:

$$\max_{\pi_1'} V_1^{\pi_1',\pi_2} + \max_{\pi_2'} V_2^{\pi_1,\pi_2'}.$$

この値が小さいほど $\boldsymbol{\pi}$ がナッシュ均衡に近いことを示す．

3 オンライン学習アルゴリズム

二人零和ゲームの均衡学習はオンライン学習アルゴリズムを用いて行うことができる．オンライン学習では，各時刻 $t = 1, ..., T$ において，学習者（プレイヤ）は凸集合 $X \subseteq \mathbb{R}^n$ から戦略 x^t を選択し，敵対者は線形損失ベクトル $\ell^t \in \mathbb{R}^n$ を選択する．本研究では，$X = \Delta^n$ と仮定する．プレイヤは，各時刻 t において x^t を選択した後に，損失ベクトル ℓ^t に関する情報を観測し，利得 $-\langle x^t, \ell^t\rangle$ を得る (損失 $\langle x^t, \ell^t\rangle$ を被る)．プレイヤの目的は累積利得 $\sum_{t=1}^T -\langle x^t, \ell^t\rangle$ を最大化することである．ここで，プレイヤが観測可能な損失ベクトル ℓ^t の情報の扱いは様々な設定が考えられている．典型的に扱われる完全フィードバック設定では，プレイヤはベクトル ℓ^t の要素をすべて観測可能である．一方，部分的フィードバック設定では，プレイヤは ℓ^t の一つの要素のみが観測できる．二人零和ゲームの均衡学習をオンライン学習の枠組みで行う場合，あるプレイヤ i の戦略は $x^t = \pi_i^t$ で，損失ベクトルは $\ell^t = Q_i^{\boldsymbol{\pi}^t}$ で与えられる．

次にオンライン学習アルゴリズムを概説する．FTRL は時刻 t におけるプレイヤの戦略 x^t を次のように定義する:

$$x^t = \arg\min_{x\in X} \left\{ \eta \left\langle x, \sum_{k=1}^{t-1} \ell^k \right\rangle + \phi(x) \right\}. \tag{1}$$

ここで η は学習率，$\phi(x) : X \to \mathbb{R}$ は正則化関数である．正則化関数を $\phi(x) = \sum_{i=1}^n x_i \log x_i$ とすると，連続時間における FTRL のダイナミクスがレプリータダイナミクスと一致す

* 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

	R	P	S
R	0	-1	3
P	1	0	-1
S	-3	1	0

表 1: Biased RPS の利得

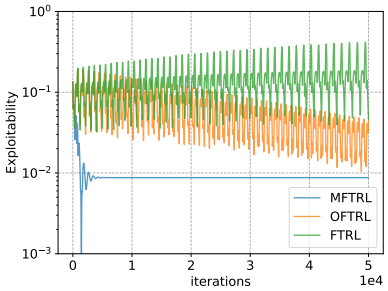

図 1: 完全情報フィードバック設定における exploitability

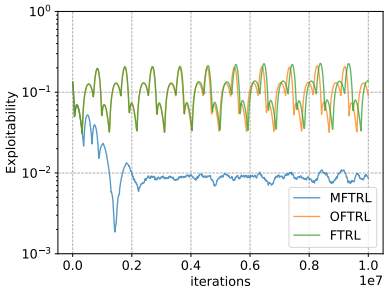

図 2: 部分的フィードバック設定における exploitability

ることが知られている [5].

OFTRL は，式 1 における累積損失に，時刻 t での損失ベクトルの予測ベクトル m^t を足し込むだけの単純な手法である:

$$x^t = \arg\min_{x \in X} \left\{ \eta \left\langle x, \sum_{k=1}^{t-1} \ell^k + m^t \right\rangle + \phi(x) \right\}.$$

本研究では，完全情報フィードバック設定では $m^t = \ell^{t-1}$，部分的フィードバック設定では [2] における方法を用いて m^t を更新するものとする．

本研究では，突然変異に着想を得た手法 Mutant Follow the Regularized Leader (MFTRL) を提案する．MFTRL は戦略 x^t を次のように計算する：

$$x^t = \arg\min_{x \in X} \left\{ \eta \left\langle x, \sum_{k=1}^{t-1} \left(\ell^k + \mu M^t \right) \right\rangle + \phi(x) \right\}.$$

ここで μ は突然変異率，$M^t = \left(\frac{1}{x_j^t} \left(x_j^t - \frac{1}{n} \right) \right)_{j \in [n]}$ とする．正則化関数を $\phi(x) = \sum_{i=1}^n x_i \log x_i$ とすると，連続時間における FTRL のダイナミクスから突然変異付きレプリータダイナミクス [6] を得ることができる．

4 計算機実験

標準形の二人零和ゲームにおいて，FTRL, OFTRL, MFTRL の 3 つの手法がどのような戦略を獲得するかを確認する．プレイヤ 1 の利得は表 1 の Biased RPS (Rock-Paper-Scissors) に従い，行がプレイヤ 1 の行動，列がプレイヤ 2 の行動に対応する．完全情報フィードバック設定と部分的フィードバック設定それぞれにおける戦略 $\boldsymbol{\pi}^t$ の exploitability を比較する．完全情報フィードバック設定では学習率 $\eta = 0.1$，突然変異率 $\mu = 0.01$ とし，部分的フィードバック設定では学習率 $\eta = 0.0001$，突然変異率 $\mu = 0.01$ とする．部分的フィードバック設定では，時刻 t の利得ベクトル $-\ell^t = Q_i^{\boldsymbol{\pi}^t}$ は，利得 $r_i^t(a)$ から重点サンプリングを用いて次のように推定する:

$$\hat{Q}_i^{\boldsymbol{\pi}^t}(a) = \begin{cases} \frac{r_i^t(a)}{\pi_i^t(a)} & \text{if } a = a_i \\ 0 & \text{otherwise,} \end{cases}.$$

図 1 に完全情報フィードバック設定における exploitability の推移を示す．最終的な exploitability は，FTRL が 0.14 と 0.035 の間で，OFTRL が 0.046 と 0.010 の間で振動した．MFTRL は exploitability が 0.0088 に収束した．exploitability の推移をみると，FTRL は学習の初めから変わらず均衡戦略に収束せずで振動した．一方，OFTRL は振動しながらゆっくりと均衡戦略に近づいていることがわかる．MFTRL は時刻 5,000 を境に，均衡戦略に近いところに収束した．

図 2 に部分的フィードバック設定における exploitability の推移を示す．最終的な exploitability は，FTRL が 0.21 と 0.035 の間で，OFTRL が 0.20 と 0.033 の間で振動した．MFTRL は exploitability が 0.0090 に収束した．exploitability の推移をみると，FTRL と OFTRL は学習の初めから変わらず均衡戦略に収束せずで振動している．一方，MFTRL は時刻 2,500,000 を境に，均衡戦略に近いところに収束した．以上から，MFTRL は完全情報フィードバックだけでなく部分的フィードバックにおいて，学習戦略が均衡に収束することを実験的に示せた．

参考文献

[1] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pp. 6629–6640, 2017.

[2] S. Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pp. 2552–2583. PMLR, 2021.

[3] M. Johanson, K. Waugh, M. H. Bowling, and M. Zinkevich. Accelerating best response calculation in large extensive games. In *IJCAI*, pp. 258–265, 2011.

[4] C. Lee, H. Luo, C. Wei, and M. Zhang. Linear last-iterate convergence for matrix games and stochastic games. *CoRR*, abs/2006.09517, 2020.

[5] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *ACM-SIAM Symposium on Discrete Algorithms*, pp. 2703–2717, 2018.

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