日本オペレーションズ・リサーチ学会
2019年 秋季研究発表会

見間違え付き繰り返しゲームにおける協力的均衡とダイナミクス

電気通信大学　*西野上和真　NISHINOUE Kazuma
05000493　電気通信大学　岩崎敦　IWASAKI Atsushi

1. はじめに

本稿では，見間違えのある繰り返しゲームにおける戦略のダイナミクスを突然変異付きレプリケータダイナミクスを用いて分析する．見間違えのある (不完全私的観測付き) 繰り返しゲームは，プレイヤが相手の行動についてノイズを含むシグナルを観測し，そのシグナルを他のプレイヤは観測できないという特徴をもつ．従来有効であると言われていたしっぺ返し戦略 (Tit-for-tat, TFT) は協力関係を維持できない．一方で，文献 [1] は協力関係に戻りやすい 1 期相互処罰戦略 (1-period mutual punishment, 1MP) という戦略が均衡となることを発見している．しかし，ある戦略がもっとも高い利得を実現する均衡だからといって，その戦略がダイナミクスの帰結で生き残るとは限らない．そこで，本稿では均衡戦略とダイナミクス [2] の帰結で生き残る戦略の関係性を吟味し，1 期片側処罰戦略 (1-period unilateral punishment, 1UP) が協力的なダイナミクスを実現するための触媒の役割を果たすことを明らかにした．

2. モデル

本章では文献 [1] に基づいて，2 人対称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 $t = 0,1,2,\ldots$ 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．成分ゲームは無限回繰り返し行われるので，プレイヤ i の割引利得和は割引因子 $\delta \in (0,1)$ により $\sum_{t=1}^{\infty} \delta^t g_i(\boldsymbol{a}^t)$ となる．$g_i(\cdot)$ は表 1 に示す囚人のジレンマの利得表に従う．

次にプレイヤ 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 を選択した際の私的シグナルは b となる．プレイヤ 2 についても同様である．シグナル分布 $o(\boldsymbol{w} \mid \boldsymbol{a})$ を両プレイヤが正しいシグナルを観測する確率は p，片方のプレイヤが間違ったシグナルを観測する確率はそれぞれ q と仮定する．また，$1-p-2q$ の確率で両方のプレイヤが間違ったシグナルを観測する．例として, プレイヤが (C, C) を選んだ後のシグナル分布を表 2 に示す．この観測構造はほぼ完全観測 (Nearly-Perfect, または Action Misperception) と呼ばれる．本稿では AM と略記する．

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

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

表 2: (C, C) のときのシグナル分布 (AM)

	$w_2 = g$	$w_2 = b$
$w_1 = g$	p	q
$w_1 = b$	q	$1-p-2q$

プレイヤの戦略は，そのプレイヤの過去の行動と受け取ったシグナルから現在の行動への写像で表現される．FSA は繰り返しゲームの戦略を簡略に表記する方法であり，本稿では，状態数 2 以下の非同相な 26 個の FSA を戦略として用いる．ここでは，状態 R で C を，状態 P で D を選ぶとする．例えば，受け取ったシグナルによらずに必ず D を選ぶ AllD，b を観測するまでは状態 R にいるが，一度でも b を観測した後は状態 P に遷移し D を選び続ける GRIM がある．また，他の重要な戦略に 1MP と 1UP がある．1MP は状態 R から開始し，g を観測している間は遷移せずに同じ状態に留まるが，b を観測したら異なる状態へと遷移する戦略である．一方で，1UP は状態 R から開始し b を観測した直後のみ D を選び，それ以外では C を選ぶ戦略である．

（突然変異付き）レプリケータダイナミクスは，ある戦略の集団において，“人口” の利得に応じて各戦略が増減するダイナミクスである [3]．ここで，戦略の集団 $\vec{x}$ の中で戦略 j が占める割合を x_j とし，$\vec{x}$ に対して戦略 j が得る利得を $f_j(\vec{x})$ とする．また，$\sum_{j=1}^{n} q_{ij} = 1$ を満たすような q_{ij} を戦略 i の子孫が戦略 j に突然変異する確率とする．このレプリケータ方程式は

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j q_{ji} - x_i \phi,\ i = 1, ..., n$$

となる．ここで $\phi(\cdot)$ を全ての戦略の利得の平均 $\sum_j x_j f_j(\vec{x})$，$f_j(\cdot)$ を $\sum_m x_j a_{jm}$ とする．ただし，a_{jm} は戦略 m を取るプレイヤと対戦する戦略 j をとるプレイヤの割引利得和である．

3. 26 戦略のダイナミクスと均衡

図 1 に，利得パラメータ g, l に対して，ダイナミクスの収束時にどの戦略がもっとも多い人口を獲得するかを示す．g, l を 0.05 ずつ [0.05, 3.00] の範囲で変化させる一方で，割引因子 $\delta = 0.9$ およびシグナル分布を $p = 0.95$, $q = 0.01$ に固定する．戦略の初期人口は一様に分布しているとし，突然変異が起きる確率を $\sum_{i \neq j} q_{ij} = 0.01$ とする．また，戦略 i が異なる戦略 j に突然変異する確率も一様とし，それぞれ $0.01/(26-1) = 0.0004$ となる．

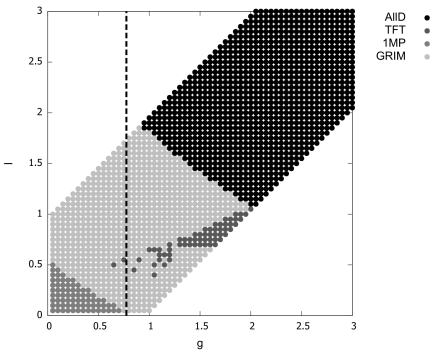


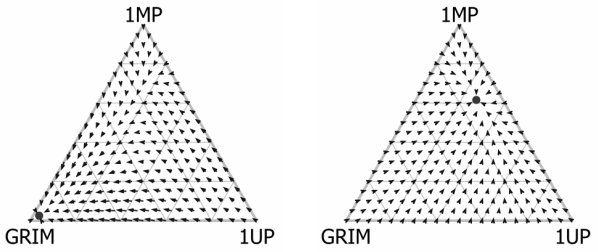
図 1: $p = 0.95, q = 0.01$ の最大人口戦略

図 1 において，g と l が十分大きいときは AllD が，十分小さいときは 1MP が生き残る．その中間領域では基本的に GRIM が生き残りやすいが，$g > l$ の範囲で TFT が生き残ることがある．TFT が生き残る領域の周辺では，TFT, GRIM, 1UP の 3 つの戦略が特定の人口比で共存することが観察された．このとき l が小さいほど，1UP の GRIM への支配力が大きくなる，つまり GRIM から 1UP へ逸脱したときの利得が大きくなるので，GRIM は生き残るのを妨げられた．その結果，どんな相手にも大きな損失がない TFT が生き残るようになった．

次に，1MP は $g \leq 0.75$ (図 1 の点線の左側) のとき，（プレイヤの利得の和に関して）最適な均衡を構成し，その 1 人あたりの利得は 9.45 となる．しかし，利得 6.93 しか実現しない均衡である GRIM の方が生き残る領域が大きい．

ここで TFT, GRIM, 1UP の 3 戦略が共存した場合と同様，1UP が重要な役割を果たすことを述べる．まず GRIM が生き残る領域から $g = 0.5, l = 1.0$ を考え，相手が GRIM をプレイすると仮定する．このとき GRIM をプレイして得られる利得は 6.93, 1MP をプレイして得られる利得は 5.54 になるので，GRIM から 1MP に逸脱するインセンティブは存在しない．また，1UP をプレイして得られる利得は 5.48 であり，やはり GRIM から逸脱するインセンティブは存在しない．この傾向は図 2a からも観察される．これらの 3 戦略しか存在しない場合，どの戦略の分布から始めても GRIM の人口が最も多くなる．さらに 26 戦略のダイナミクスでも，早い段階でこれら 3 つ以外の戦略がほとんどなくなり，GRIM に収束することを確認した．

次に g を 0.5 に固定したまま，l を 0.1 に下げた場合を考える．相手が GRIM をプレイするとき，GRIM は 6.99, 1MP は 6.86, 1UP は 6.85 の利得を与える．相手が 1UP をプレイするとき，GRIM は 9.19, 1MP は 9.59, 1UP は 9.42 の利得を与える．利得表上では 1UP は GRIM を支配している訳ではないが，相手が一定の確率でこれら 3 つの戦略をプレイするとき，自分は 1UP や 1MP をプレイする方が期待利得が高くなる．実際，図 2b では GRIM に収束することはなくなり，3 つの戦略が共存するようになる．これは裏切られたときの損失 l が小さくなることで，相手を絶対に許さない GRIM より，裏切られて少しばかり損をしても将来の協力を見越して協力に戻る 1UP や 1MP が生き残るようになったと考えられる．



(a) $g = 0.5, l = 1.0$　(b) $g = 0.5, l = 0.1$

図 2: GRIM-1MP-1UP 間のダイナミクス

参考文献

[1] ジョヨンジュン, 岩崎敦, 神取道宏, 小原一郎, 横尾真. 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.

[2] Drew Fudenberg Lorens A. Imhof and Martin A. Nowak. Evolutionary cycles of cooperation and defection. *in Proceedings of the National Academy of Sciences*, Vol. 102, No. 31, pp. 10797–10800, 2005.

[3] Karl Sigmund. *The Calculus of Selfishness*. Princeton University Press, 2010.