5B-03

取り違えのある繰り返し囚人のジレンマにおける単独裏切-相互同期戦略

村井　伸一郎 *　　岩崎 敦 *
Shinnchiro Murai　　Atsushi Iwasaki

1　はじめに

繰り返しゲームは，長期的関係にあるプレイヤ間の（暗黙の）協調を説明するためのモデル [1] であり，主に経済学分野で企業間の談合といった協調行動を分析するために発展してきた．2 人がまったく行動を取り違えないならば，常に裏切り（ALLD）や一度でも裏切られたら許さない（Grim-trigger, GRIM）といった戦略しか生き残らないことが知られている [2]．しかし，実際の人間はしばしば行動を取り違えることがある．例えば，協力しようとしたが失敗してしまったり，サボったつもりがうまくいってしまったりするのは自然である．こうした行動の取り違えは進化ゲーム理論における重要な仮定であり，実際，こうした間違いがないと，お互いに協力することが進化的安定性を満たさないことが知られている．

本研究では，プレイヤたちが一定の確率で意図した行動と異なる行動を取ってしまう，行動の取り違え (implementation errors) [3] が発生するとき，突然変異付きレプリケータダイナミクスの帰結がどうなるかを吟味する．行動の取り違えについては広範な先行研究があるが，その多くは戦略空間をかなり限定する，もしくは戦略自体を進化させるような閉じていない戦略空間を想定している．もっともよく使われる戦略空間として，一期記憶戦略 (memory-one strategies) がある [4]．これは，今日の自分の行動を，昨日の自分と相手の行動から決める戦略のクラスであるが，自分が行動を取り違えたかどうかを考慮していなかった．

そこで本研究では，戦略を有限状態機械 (Finite State Automaton, FSA) [2, 5] を用いて，自分の意図した行動と実現した行動を区別した 482 個から成る戦略空間におけるダイナミクスの帰結を分析した．その結果，割引因子が十分に小さく，行動の取り違えが起こりにくいとき，従来の一期記憶戦略において有名な “勝ち残り，負け逃げ” (Win-Stay, Lose-Shift, WSLS) よりも “単独裏切，相互裏切” (Unilateral Defection, Mutual Defection, UDMD) という戦略が生き残ることを世界で初めて発見した（図 1a と 1b）．

2　モデル

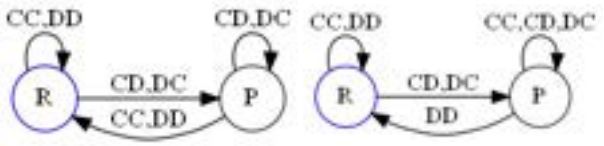
本章では行動の取り違えのある無限回繰り返しゲームをモデル化する．ここでプレイヤ $i \in \{1,2\}$ はステージゲームを無限期間 $t = 0,1,2,\cdots$ に渡って繰り返す．割引因子は $\delta \in (0,\ 1)$ とする．各期においてプレイヤ i は有限集合 $A_i = \{C,\ D\}$ から行動 a_i を選択し，その行動の組を $\boldsymbol{a} = (a_1, a_2) \in A^2$ とする．このとき，意図した行動を $\bar{a}$，実現した行動を $\hat{a}$ とする．プレイヤの利得は表 1 に示す囚人のジレンマの利得表に従う．また，意図した行動の組 $(\bar{a}_1, \bar{a}_2)$ に対して，実現した行動の組 $(\hat{a}_1, \hat{a}_2)$ が生起する同時確率をエラー分布 $o((\hat{a}_1, \hat{a}_2)|(\bar{a}_1, \bar{a}_2))$ とする．

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

	$\hat{a}_2 = C$	$\hat{a}_2 = D$
$\hat{a}_1 = C$	$1,\ 1$	$-l,\ 1+g$
$\hat{a}_1 = D$	$1+g,\ -l$	$0,\ 0$

表 2:　エラー分布 $o((\hat{a}_1, \hat{a}_2)|(\bar{a}_1, \bar{a}_2))$

	$\hat{a}_2 = \bar{a}_2$	$\hat{a}_2 \neq \bar{a}_2$
$\hat{a}_1 = \bar{a}_1$	p	q
$\hat{a}_1 \neq \bar{a}_1$	q	$1-p-2q$



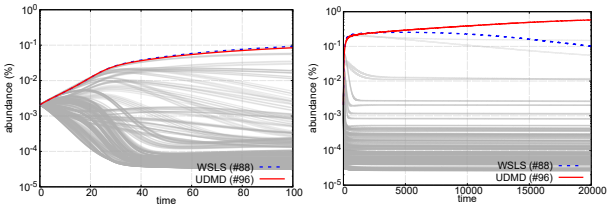
(a) WSLS (#88)　　(b) UDMD (#96)

図 1: WSLS と UDMD

繰り返しゲームの戦略は，昨日までの履歴から今日の選択する行動への写像で定義する．本研究では行動を取り違えた後の振る舞いを網羅し，状態数 2 以下の非同相な 482 個の FSA を戦略空間とする．FSA の状態は R (reward, 報酬) と P (punishment, 処罰) の 2 つに区別され, プレイヤ i は状態 R で行動 $a_i = C$ を選び, 状態 P で行動 $a_i = D$ を選ぶ．それぞれの状態でプレイヤは自分と相手がとった行動で次にどの状態に遷移するかが決まる．例えば，状態 R からは 4 つの行動の組に対して状態遷移が決まる．図 1 の各 FSA において，状態 R における CC や CD は自分が行動を取り違えなかったときの，DC や DD は自分が行動を取り違えたときの遷移を表す．

図 1a に WSLS を示す．ここで#88 とは, FSA 戦略を列挙する上での番号を表す．プレイヤが WSLS にしたがうとき，最初は協力 (状態 R) し, 自分もしくは相手のどちらかが裏切るまで協力する．どちらかが裏切った後は, お互いに裏切るもしくは協力して初めて協力に戻る．この戦略は一期記憶戦略に属する, つまり, どちらの状態にいても, 実現した行動の組に対する状態遷移が共通する戦略となっている．協力のコストが十分小さいとき, WSLS は一期記憶戦略空間で, 最も生き残

* 電気通信大学大学院情報理工学研究科



(a) 100 期まで　(b) 20000 期まで

図 2: $g = l = 0.1$ におけるダイナミクス

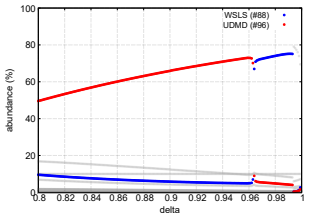


図 3: 割引因子 δ の影響

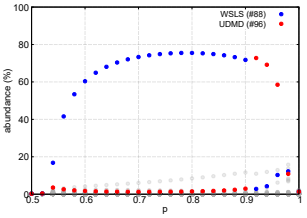


図 4: 取り違えない確率 p

りやすくなる．一方で, 482 個の戦略空間において, 図 1b に示す UDMD (#96) が WSLS 以上に生き残りやすいことがわかった．UDMD は WSLS と似ているが, 状態 P でお互いに協力できても, 状態 R に戻らない戦略になっている．相互協力の後の状態遷移が状態 R からと状態 P からとで異なる点で, 一期記憶戦略に属さない.

このような数ある戦略の中から有効な戦略を発見する方法の 1 つとして，突然変異付きレプリケータダイナミクス [5] がある．本論文では，その方程式を

$$\dot{x_i} = x_i \left[f_i(\vec{x}) - \phi(\vec{x}) \right] + u \left(\frac{1}{n} - x_i \right), \ i = 1, ..., n \quad (1)$$

と定義する．$\phi(\cdot)$ を全ての戦略の利得の平均 $\sum_j x_j f_j(\vec{x})$, $f_j(\cdot)$ を $\sum_m x_m a_{jm}$ とする．ただし，a_{jm} は戦略 j をとるプレイヤが戦略 m を取るプレイヤと無限回プレイしたときの割引利得和である.

数値実験では, 50000 期の帰結を分析した．エラー分布 $o((\hat{a}_1, \hat{a}_2)|(a_1, a_2))$ は表 2 に従う．割引因子 δ を 0.90, 突然変異率 u を 0.01 とした．g と l は $[0.1, 3.0]$ の範囲で 0.1 刻みで変化させ, $|g - l| < 1$ となる組のみを使用した.

3　取り違えがある環境下のダイナミクス

$g = l = 0.1$ における戦略の人口比率の時間変化を図 2a に 100 期まで, 図 2b に 20000 期までを示す．エラー分布のパラメータを $p = 0.95$ および $q = 0.01$ とし, 初期の戦略は一様に分布すると仮定した．まず, WSLS および UDMD はそれぞれ $1/482 \approx 0.002$ だけ存在し, 100 期までにそれぞれ $(0.092, 0.084)$ にまで増加し, WSLS は最大多数戦略となる．最大多数戦略とは, 収束時に最も多くの人口を獲得した戦略を意味する．さらにおよそ 2300 期で, 両戦略は比率の差が広がり始め, 20000 期以降でほぼ収束し, その割合は $(0.058, 0.660)$ になり, UDMD が WSLS を駆逐することになる．しかし, UDMD は常に WSLS を駆逐するとは限らない．実際, 割引因子が十分高いケースや行動を取り違える確率が十分小さいときは WSLS が最大多数を占めるようになる.

図 3 に, 割引因子に対してダイナミクスが収束したときの戦略の分布がどのように変化するかを示す．横軸が $[0.800, 0.999]$ の範囲で 0.001 刻みで動かした割引因子 δ を表し, 縦軸がダイナミクス収束時の戦略の比率を表す．その他のパラメータは $p = 0.95$, $q = 0.01$, $g = l = 0.1$ とした．割引因子とはプレイヤが将来利得をどれだけ重要視するかを表す．割引因子 δ が 0.96 以下のとき，UDMD が最大多数になり, 0.96 を越えると WSLS が最大多数になる．これは, 割引因子が大きくなることで, プレイヤが将来利得をより重視した結果, UDMD より裏切りの後に相互協調を回復させやすい WSLS が生き残るようになったと考えられる．実際，UDMD が状態 P から状態 R に戻るのは, DD が実現した後だけだが, WSLS は CC が実現した後でも状態 R に戻るようになっている分相互協調を回復させやすい.

次に，エラー分布のパラメータがダイナミクスの帰結に与える影響を分析するために，図 4 に 2 人のプレイヤが行動を取り違えない確率を表す p に対する収束時の戦略比率の変化を示す．具体的には，$g = 0.1, l = 0.1, q = 0.01, \delta = 0.90$ に対して p を $[0.50, 1.00]$ の範囲で 0.02 刻みで変化させたときの 482 戦略の比率を示す．UDMD と WSLS 以外の戦略は灰色の線としている．もし行動の取り違えが起きるのが 100 回のプレイのうち 10 回未満 ($p \geq 0.92$) であれば UDMD が, 10 回以上 ($p < 0.92$) であれば WSLS が, 収束時に最大多数となる．ここではいずれかが行動を取り違える確率 q を固定しているため, p が十分小さくなると, 相互処罰の状態から相互協調に戻る確率が UDMD より WSLS の方が相対的に大きくなり，p の大きさによって最大多数戦略が不連続的に変化する.

参考文献

[1] 神取道宏. 人はなぜ協調するのか –くり返しゲーム理論入門–. 三菱経済研究所, 2015.

[2] 西野上和真, 五十嵐瞭平, 岩崎敦. 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. 情報処理学会論文誌, Vol. 63, No. 4, pp. 1138–1148, 2022.

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

[4] Christian Hilbe, Krishnendu Chatterjee, and Martin Nowak. Partners and rivals in direct reciprocity. *Nat Hum Behav 2*, p. 469–477, 2018.

[5] Benjamin Zagorsky, Johannes Reiter, Krishnendu Chatterjee, and Martin Nowak. Forgiver triumphs in alternating prisoner's dilemma. *PLOS ONE*, pp. 1–8, 2013.