情報処理学会論文誌 Vol.63 No.4 1138–1148 (Apr. 2022)

推薦論文

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

西野上 和真[1,a)] 五十嵐 瞭平[2,b)] 岩崎 敦[2,c)]

受付日 2021年7月6日, 採録日 2022年1月11日

概要：本論文は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を分析した．私的観測は，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持つ．ここで，どんな戦略の組が均衡になるかはゲーム理論の有名な未解決問題の1つであり，本論文では戦略空間を状態数2以下の有限状態機械に限定したレプリケータダイナミクスの帰結から，どのような戦略が生き残るかを吟味した．その結果，利得構造に応じて，4つの社会（非協力，不寛容，相互協力，周期協力）が現れることが分かった．とくに周期協力社会で，非専門家の間では有効と信じられてきたしっぺ返し戦略は最大多数になりうるが，他の戦略と共存しなければならないという不安定さを持つ．一方，他の社会では，ある特定の戦略が人口のほぼすべてを占めるようになる．さらにノイズと突然変異率に関する感度分析から十分広いパラメータにおいて同じ傾向を保つことが分かった．

キーワード：ゲーム理論，繰り返しゲーム，囚人のジレンマ，レプリケータダイナミクス

Dynamics of Cooperation in Repeated Games with Private Monitoring

Kazuma Nishinoue[1,a)] Ryohei Igarashi[2,b)] Atsushi Iwasaki[2,c)]

Received: July 6, 2021, Accepted: January 11, 2022

Abstract: This paper analyzes the dynamics of cooperation in a repeated prisoner's dilemma under private monitoring, where each player privately observes noisy signals about the opponent's actions. What kind of strategy forms an equilibrium is one of the open, but fundamental questions in game theory. We examine what kind of strategies are abundant in the consequences of replicator-mutator dynamics, where the strategy space is restricted to a finite state automata within two states. As a result, we found that four kinds of societies (non-cooperation, intolerant cooperation, mutual cooperation, and cyclical cooperation) emerge according to payoff structures. In particular, in the cyclical cooperation society, the well-known Tit-for-tat strategy, which is believed to be successful for non-experts, can become the most abundant. However, it is not so stable because it must coexist with other strategies. In the other societies, a particular strategy dominates the population. Furthermore, our sensitivity analysis on the parameters of signal distributions and mutation rate shows that the trend persists for a wide range of parameters.

Keywords: game theory, repeated games, prisoner's dilemma, replicator dynamics

1. はじめに

無限回繰り返しゲームは，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である [17]．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

1 SCSK 株式会社
SCSK Corporation, Koto, Tokyo 135–8383, Japan
2 電気通信大学大学院情報理工学研究科
Graduate School of Informatics and Engineering, The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
a) n1930095@edu.cc.uec.ac.jp
b) i2030011@edu.cc.uec.ac.jp
c) atsushi.iwasaki@uec.ac.jp

本論文の内容は 2020 年 9 月の FIT2020 第 19 回情報科学技術フォーラムにて報告され，同プログラム委員長により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

ために発展してきた [20]．暗黙の協調を実現するには，プレイヤが相手の行動をある程度観測できることが前提となる．これまで，相手の行動が完全に観測できる完全観測 (perfect monitoring) のケースについては多く論じられている [1], [11], [14]．しかし，現実には相手の行動が完全に観測できない不完全観測 (imperfect monitoring) のケース，つまり，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場合がある．これはとくに，不完全私的観測 (imperfect private monitoring) のケースと呼ばれる [6], [18], [21]．不完全私的観測付き無限回繰り返しゲーム (infinite repeated games with imperfect private monitoring) の特徴は，プレイヤが相手の行動に関してノイズを含む観測（シグナル）を私的に受け取ると仮定する点にある．いいかえると，あるプレイヤが相手の行動について観測したシグナルと異なるシグナルを他のプレイヤが観測しているかもしれない．不完全私的観測付き無限回繰り返しゲームおいてどのような振舞い（戦略）が最適なのかについては，ゲーム理論における代表的なゲームである囚人のジレンマの例でさえ十分に分かっていない．たとえば，部分観測可能マルコフ決定過程 (Partially Observable Markov Decision Process, POMDP) を用いて均衡を計算する手法 [17] が知られているが，その計算量は一般には決定不能と知られている．

そこで本論文では，均衡の代わりに突然変異付きのレプリケータダイナミクス [3], [15] の帰結を用いて，私的観測下の繰り返し囚人のジレンマでどんな戦略が生き残るかを分析する．レプリケータダイナミクスは，進化ゲーム理論でよく用いられるダイナミクスの 1 つであり，頻度依存淘汰モデルを用いて最適な戦略を探るため，その帰結が比較的計算しやすい．無限に大きな集団を仮定し，各戦略をとるプレイヤの頻度の時間的変化を計算する．無限集団を仮定することでモデルの持つ確率性を無視できるので，そのダイナミクスは決定論的となり，微分方程式で記述できる [19]．利得が高くなる戦略をとるプレイヤの人口は増加し，低くなる戦略をとる人口はより良い戦略へとって代わられてやがて絶滅するといった具合に自然淘汰の過程を表現する．厳密には，均衡とダイナミクスの帰結の 2 つに包含関係はない．ある戦略の組が均衡になったとしても，それがダイナミクスの帰結で最大多数を占めるとは限らない．その逆も必ずしもいえない．また，均衡は複数存在しうるので，ダイナミクスがどの均衡に収束するかを事前に予測することは困難であり，そもそも均衡を構成する戦略が存在しない場合もある．そのような場合でも，自然淘汰のダイナミクスはどんな戦略が生き残るかを示せる．

理論生物学や進化ゲームの文脈では，自分の行動を出し間違える摂動 (trembling-hand) はさかんに研究されてきた [14]．しかし，その重要性にもかかわらず，相手の行動を見間違える私的観測を進化ゲームの文脈で網羅的に分析することは非常に難しいと考えられてきた．その理由の 1 つとして，摂動と私的観測は戦略と情報の構造が異なるため，従来の成果が適用できないことにあげられる．私的観測では相手に誤解を与えてしまったかどうかを知ることはできないのに対し，摂動では相手に誤解を与えてしまったか自分で分かる，すなわち自分と相手がとった行動をお互いに観測できる．そのため，2 人が実際にとった行動の組を（公的な）シグナルと解釈した公的観測の問題と摂動の問題は等価になる．公的観測とは，プレイヤがお互いの観測を共有する不完全観測のクラスであり，完全観測とほぼ同じ性質をもつことが知られている [9]．このため，摂動の場合は完全観測と同じような結果が得られると予想できる．

また，一般には複雑な行動計画となる無限回繰り返しゲームの戦略を有限状態機械 (Finite State Automaton, FSA) で記述するとき，私的観測をどのようにモデル化し期待利得を計算するかよく分かっていなかった．そこで本論文では，まずプレイヤがとりうる戦略を状態数 2 以下の FSA に限定する．つまり，プレイヤの今日とった行動と観測したシグナルから明日の行動への写像を考える．摂動の場合はこれに加えて，自分が行動をとり間違えた後の振舞いを考慮しなければならなくなる．このため，たとえば私的観測における GRIM と摂動における GRIM は一意に対応させることができない．そこで本論文では，摂動との正確な対応づけや比較は今後の課題としつつ，状態数 2 以下の非同相な FSA を列挙した．それらの期待利得はマルコフ決定過程に基づいて計算し，その利得表をもとに突然変異付きレプリケータダイナミクス [4] を計算する．

その結果，利得構造に応じて，4 つの社会（非協力，不寛容，相互協力，周期協力）が現れることが分かった．非協力社会では，つねに裏切る戦略 (ALLD) のみが生き残る，つまり単独の戦略で人口のほぼすべてを占める．次に不寛容社会ではトリガー (Grim trigger, GRIM) 戦略と呼ばれる，はじめに協力し相手が一度でも裏切ったら二度と許さない戦略が生き残る．さらに相互協力社会では，1 期相互処罰 (1-period Mutual Punishment, 1MP) という新しい戦略が生き残る [17]．最後に，非自明な均衡戦略がないときに発生する周期協力社会では，非専門家の間では有効と信じられてきたしっぺ返し戦略 (Tit-For-Tat, TFT) が最大多数になりうるが，他の戦略とサイクルを構成し共存しなければならず，単体の戦略として安定しないことが分かった．さらにノイズと突然変異率に関する感度分析で十分広いパラメータでも傾向が変わらないことを示した．

最後に，本論文の成果は，情報処理学会のスコープと合致しないと思う読者が多いかもしれない点について補足する．そのような読者は，もっと一般的な（いかにも現実的な状況をモデル化したかのように見える）設定における協調行動の仕組みの解明を期待するかもしれない．しかし複雑なモデル上の複雑な挙動をきちんと理解するには，シン

プルなモデル上の挙動を理解する必要がある．これに対して本論文の分析は，複雑なシミュレーションがどのような場合にうまくいくのか？ なぜうまくいくのかを理解するためのものさしという価値がある．なぜうまくいくのかを理論的に説明できないシミュレーションから役に立つ洞察を引き出すのはきわめて難しい．仮に引き出せたとしてもとても安心して一般化できないだろう[*1]．本論文はこのようなゲーム理論と（ゲーム理論を模したつもりでいる）シミュレーション研究のギャップを埋める一助になるだろう．

2. モデル

本章では文献 [17] に基づいて，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 $i \in \{1,2\}$ はステージゲームを無限期間 $t = 0, 1, 2, ...$ にわたって繰り返す．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 は自分以外のプレイヤ $-i$ の行動 $\boldsymbol{a}_{-i}$ 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．ステージ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．ただし，$g_i(\cdot)$ の値は利得表によって定められた値に従う．

ここで不完全観測付き無限回繰り返しゲームでは，ステージゲーム利得は毎期変動する利得の平均値となっている．プレイヤはステージゲーム利得を直接観察するわけではない．そうでなければ，プレイヤは自分の行動と利得から相手の行動を推測できてしまう．この毎期変動する利得 π_i を，各期の自身の行動 a_i と相手の行動を示唆するシグナル ω_i から決まる実現利得と仮定し，そのシグナル分布に関する期待値をステージゲーム利得とする．本論文では，とくに断りがない限り，ステージゲーム利得を「利得」と表記する．

本論文では利得表として**表 1** に示す囚人のジレンマを用いる．表中の C は協力行為を，D は裏切り行為を表す．囚人のジレンマの利得構造は $g > 0$，$l > 0$ であり，このとき D は厳密な支配戦略となる．また，囚人のジレンマでは $|g - l| < 1$ が要求される．もしこの条件が成り立たないとすると，繰り返し囚人のジレンマにおいて協力と裏切りを交互に出すほうが，純粋な協力よりも利得が高くなってしまい，純粋な協力が維持できなくなる．

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，$\omega_1 \in \{g, b\}$ (*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．よく使われる不完全私的観測のシグナル分布にほぼ完全観測がある．ここでは，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，(C, C) が実現した場合のシグナル分布を**表 2** に示す．ただし，両プレイヤが正しいシグナルを観測する確率 p が最も高くなるように設定する．

表 1 囚人のジレンマ（$g > 0$, $l > 0$，および $|g - l| < 1$）
Table 1 Prisoner's dilemma ($g > 0$, $l > 0$, and $|g - l| < 1$).

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

表 2 (C, C) のときのシグナル分布
Table 2 The joint probability distribution of signals for (C, C).

	$w_2 = g$	$w_2 = b$
$w_1 = g$	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．無限回繰り返しゲームの戦略は，文字どおり無限個存在し，そのすべてを網羅するのは不可能である．このため，先行研究ではプレイヤの戦略を何らかの形で制限している [10], [16]．そこで本論文では，有限状態機械 (Finite State Automaton, FSA) による戦略表記を採用し，レプリケータダイナミクスで計算可能な戦略空間を定義する．同相 FSA とはまったく同じ行動パターンをとる FSA のことをいい，同相 FSA も含めて列挙すると戦略の個数は，Θ を FSA の状態数として，$|A||\Theta|^{|A|}|\Theta|^{|\Theta||\Omega|}$ となる．これに対して，同相な FSA をまとめると状態数が 2 の場合は 26 個，状態数が 3 の場合は 1,054 個の非同相な FSA が戦略空間を定義する．しかし，現状では 1,054 個の戦略でレプリケータダイナミクスを安定的に計算できないうえ，計算結果を解釈するのが非常に難しくなる．このため，本論文では状態数 2 以下の非同相な FSA が定義する戦略空間で何が起こるかを明らかにする．

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，レプリケータダイナミクスがある．ゲームを行うプレイヤの集団を考え，プレイヤはレプリケータ方程式によって求めた戦略の分布にしたがって他のプレイヤとゲームを行い利得を得る．その後，戦略の集団に対する利得と集団全体の平均利得との差に応じて戦略の人口比を増減させる [14]．本論文では突然変異の概念を導入したレプリケータダイナミクスを用いる．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

[*1] この点については文献 [23] の優れた論考を参照されたい．

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 となる確率とおく．このとき，突然変異付きのレプリケータ方程式は以下のように表される．

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

$\phi(\cdot)$ をすべ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 をと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，割引利得 ($\delta = 0.9$) を固定したうえで，g, l を $[0.05, 3.00]$ の範囲で 0.05 刻みで変化させた．戦略として状態数 2 以下の非同相な 26 個の FSA を用いるを用いる．付録の図 A·1 にこれら 26 個の FSA を列挙した．また，初期時点において，各戦略の人口は一様に分布，つまり，各戦略の存在比率はすべて等しいものとする．さらに，突然変異を起こす確率 $\sum_{i \neq j} q_{ij} = \mu$ を 0.01 とした．このとき，戦略 j から戦略 $i \neq j$ に突然変異する確率 q_{ji} は $\mu/(26-1)$ の等確率で起こるとする．この微分方程式を解く際は期数の刻み幅 Δt を可変とした Dormand-Prince 法 [2], [13] を採用し，その実装には scipy を用いた [5]．さらに Δt の最大値は 0.5 とした．ダイナミクスはすべての戦略の 1 期あたりの人口の変化量 $|\dot{x}|$ が 10^{-5} 以下となった時点で収束と判定した．また，50000 期までに収束と判定されなかった場合は計算を終了する．

3. 主要な戦略とナッシュ均衡

本章では，繰り返し囚人のジレンマにおいて重要な戦略とのそのナッシュ均衡を概説する．繰り返しゲームの戦略は過去の行動と観測の履歴から現在の行動への写像で定義される．一般には複雑になる戦略でも FSA を用いて簡略に表記できる．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．状態数 1 の戦略には ALLC（**図 1** (a)）と ALLD（図 1 (b)）の 2 つが存在し，ALLC は状態 R のみを持ち毎期必ず協力する戦略，ALLD は状態 P のみを持ち毎期必ず裏切る戦略である．一方で状態 R と P を持つ状態数 2 の著名な戦略としては，まず最初に無限期罰則のトリガー戦略 (Grim trigger, GRIM) があげられる（図 1 (c)）．GRIM は最初に協力し，相手の裏切りを観測するとそれ以降裏切り続ける戦略であり，多くの場合 GRIM は完全観測，不完全観測の両方の下で均衡を構成できる．別の戦略としては “しっぺ返し” (Tit-For-Tat, TFT) がある（図 1 (d)）．TFT は，状態 R からスタートし，相手の協力を観測した次の期では協力を，裏切りを観測した次の期には裏切りを行う戦略である．完全観測下では協力関係を維持できる一方で，不完全観測下では，いったん相手が裏切ったというシグナルを観測すると再び協力状態に戻るのは難しくなる．他にも重要な戦略として，“1 期相互処罰” (1 Mutual Punishment, 1MP) が存在する（図 1 (e)）．1MP [8] は，従来 “Pavlov” または “win-stay, lose-shift” [12] として知られている．1MP は状態 R からスタートし，相手の協力を観測したときは同じ状態にとどまり，裏切りを観測するともう 1 つの状態へと遷移する．裏切りを観測してから協力に戻るのは一見不自然に見えるが，お互いを処罰してから協力に戻ることで，見間違えのある環境で TFT より協力状態を維持しやすくなっている．最後に，“1 回処罰” (One-Shot Punishment, OSP，図 1 (f)) もしばしばダイナミクスに含まれる戦略である．OSP は状態 R からスタートし，相手の裏切りを観測した次の期のみ裏切る（状態 P に遷移する）が，その後は何を観測しても協力に戻る（状態 R に遷移する）戦略である．

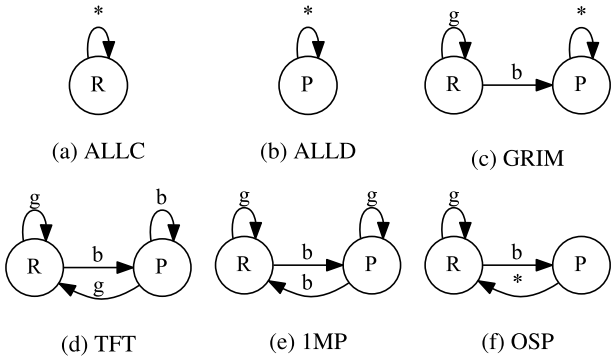


図 1 主要な FSA
Fig. 1 Representative FSAs.

次に各プレイヤの戦略空間を 26 個の FSA に限定した 2 人ゲームのナッシュ均衡を考える．相手がある FSA にしたがってプレイするとき，自分の割引利得和を最大化する FSA を最適反応 FSA と呼ぶ．ある FSA の組がナッシュ均衡になるとは，その組がお互いに最適反応となる FSA になっていることをいう [22]．完全観測の場合，割引因子 δ が十分に大きければ，ALLD や GRIM, 1MP, TFT を含む多くの戦略が均衡を構成する．不完全観測の場合，厳密な均衡条件を解析的に求めるのは難しいが，均衡を構成するのは ALLD, GRIM, 1MP および状態 P から始める 1MP の 4 種類のみである．とくに TFT が不完全観測で均衡を構成することはない [17]．たとえば，$p = 0.95$, $q = 0.01$, $\delta = 0.9$ のとき，ALLD はつねに均衡を構成し，GRIM は l がおおよそ 0.15 より大きければ均衡を構成する．状態 R もしくは P から始める 1MP は g が 0.75 より小さければ均衡を構成する．先に述べたように均衡とダイナミクスの帰結に包含関係はないが，ダイナミクスの帰結でどんな均衡戦略（もしくは均衡でない戦略）が生き残るかを知ることは協力の仕組みを理解するうえで重要である．

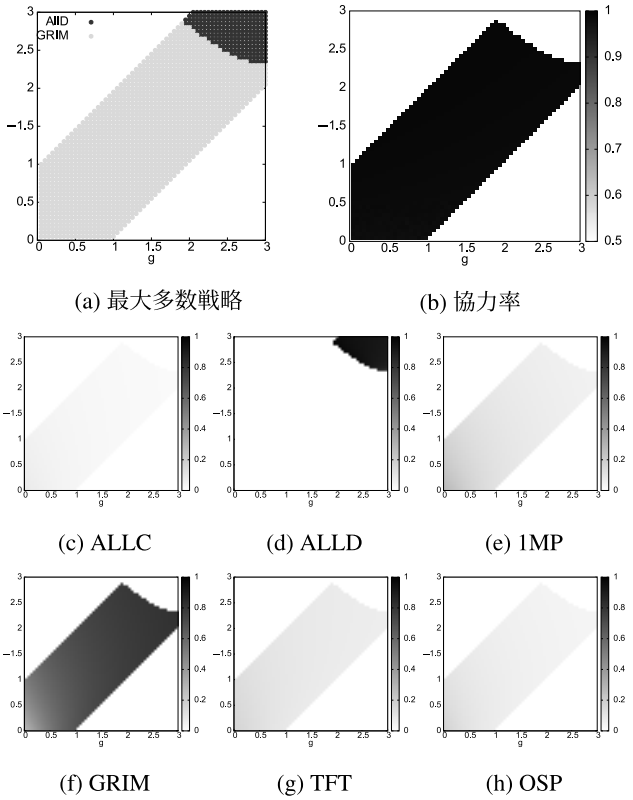


図 2 完全観測におけるレプリケータダイナミクスの帰結：$p = 1.00$, $q = 0.00$

Fig. 2 Consequences of replicator dynamics under perfect monitoring: $p = 1.00$, $q = 0.00$.

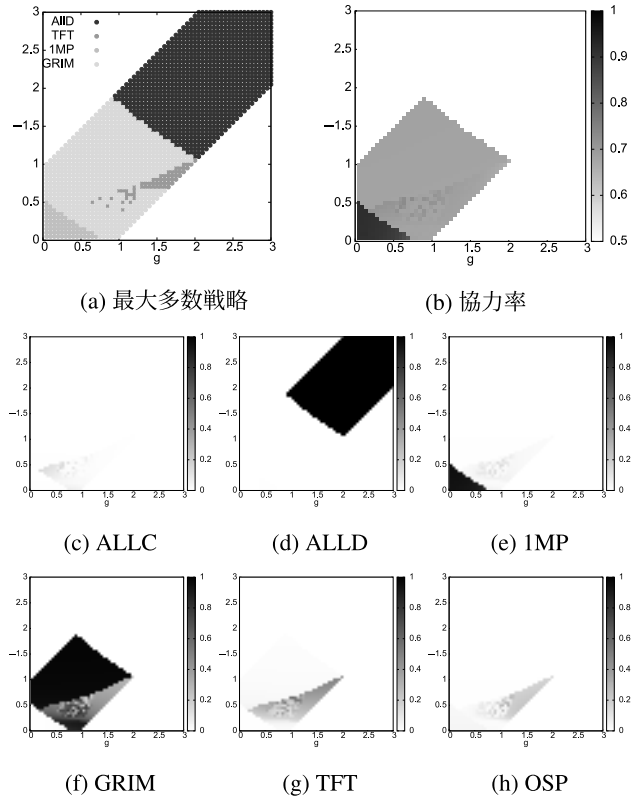


図 3 ほぼ完全観測におけるレプリケータダイナミクスの帰結：$p = 0.95$, $q = 0.01$

Fig. 3 Consequences of replicator dynamics under almost perfect monitoring: $p = 0.95$, $q = 0.01$.

4. 2 状態戦略間のダイナミクス

4.1 完全観測下におけるダイナミクス

図 2 に完全観測のダイナミクスの帰結を示す．ここで，シグナル分布パラメータを $p = 1$ および $q = 0$ とする．それぞれの図の横軸は自分の裏切りによる利得の増分 g，縦軸は相手の裏切りによる損失 l に対応し，0.05 刻みで $[0.05, 3.00]$ をプロットした．図 2 (a) に収束時に最も多くの人口を獲得した戦略，すなわち最大多数戦略を，図 2 (b) は協力率を示している．これは収束時の戦略人口比に対して無限回繰り返しゲームを行うとして実現する (C, C) の頻度から計算した値である．残りの図 2 (c)–図 2 (h) は主要 6 戦略の収束時の人口比を示している．

図 2 (a) では，g と l が十分大きい領域では ALLD が，それ以外の領域では GRIM が最大多数戦略となる．ALLD の人口比は約 9 割に到達する．一方で，GRIM が最大多数となるとき，他の 4 つの戦略 (ALLC, 1MP, TFT, OSP) とそれなりの割合で共存する．どれくらいの割合で共存するかは g, l の値に依存し，g, l が大きくなるにつれて GRIM の占める割合が増加する．図 2 (b) では，ALLD が最大多数となるときの協力率はほぼ 0 である一方，GRIM が最大多数となるときは 0.97 を上回る．これは図 2 (c)–図 2 (h) にあるように GRIM を含む 5 つの戦略はお互いに恒久的な協力関係を実現するためである．

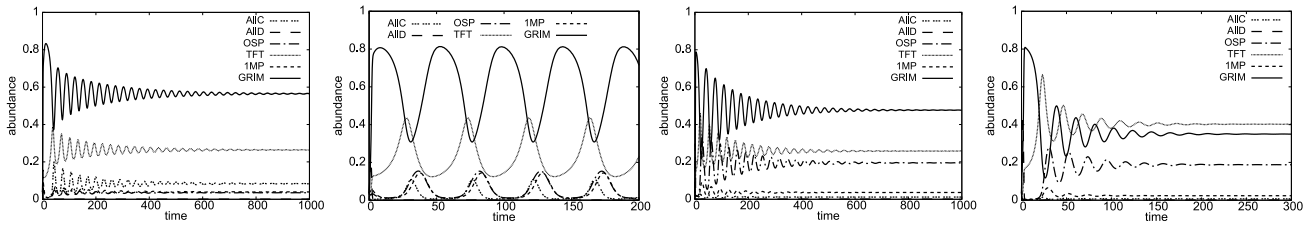
4.2 ほぼ完全観測と 4 種の社会

図 3 にほぼ完全観測のダイナミクスの帰結を示す．ここで，シグナル分布パラメータを $p = 0.95$ および $q = 0.01$ とする．完全観測のときと同様に，g および l を変化させながら図 3 (a) および図 3 (b) のそれぞれに最大多数戦略と協力率を，図 3 (c)–図 3 (h) に主要戦略の人口比率を示した．

図 3 (a) が示すように，どんな戦略が生き残るかは利得構造に依存し，おおまかに 4 つの領域に分けることができる．本論文ではこの 4 つの領域を以下の 4 つの社会に分類する．

非協力社会：ALLD が最大多数となる領域
不寛容社会：GRIM が最大多数となる領域
相互協力社会：1MP が最大多数となる領域
周期協力社会：複数の戦略が共存もしくは周期を構成する領域

非協力社会は g および l が十分大きいときに発生する．このとき，裏切る誘引や裏切られることによる損失が大きいため，他のどの戦略も協力を維持するに十分な将来利得を獲得できない．不寛容社会は g および l がそこそこの大きさでかつ，g が l より小さいときに発生する．ここでは GRIM が最大多数を占めるため，最初はお互いに協力す



(a) $g = 0.50, l = 0.40$ (b) $g = 0.80, l = 0.40$ (c) $g = 1.20, l = 0.40$ (d) $g = 1.50, l = 0.75$

図 4 複数の戦略が共存する利得パラメータにおける戦略分布のダイナミクス

Fig. 4 Population dynamics with the payoff parameters where multiple strategies gradually coexist.

るが，一度でも裏切りが発生すると，永遠に裏切り続けることになり，相手を許すことはなくなる．実際 l が十分に大きいときは，裏切られることによる損失が大きくなるため，寛容な戦略で協力を回復する誘引を提供しにくくなる．GRIM は見間違えが起こるまでは協力状態を維持できるため，図 3 (b) に示すようにその協力率は 0.680 程度となる．全体としては TFT も若干生き残るがその人口比率はわずか 0.01 程度にしかならない．

次に相互協力社会は g および l が十分小さいときに発生する．1MP どうしの対戦では，どちらか一方のプレイヤが *bad* を観測して協力状態が途切れた後も，互いに裏切り合う相互処罰を経て，協力状態に簡単に戻ることができる．互いに罰を与えることで相互協力に戻るのは，一見直感に反するが，相互処罰がうまく協力に戻るタイミングを明確にしている．そのため，1MP の協力率は見間違えのある状況でも約 0.928 と非常に高く，急速に人口を獲得する．また，図 3 (h) にあるように OSP がわずかに存在するが，その比率は 0.05 以下にしかならない．

最後に，周期協力社会は g および l がそこそこの大きさでかつ，g が l より大きいときに発生する．ここで最大多数となる戦略は GRIM もしくは TFT のいずれかになるが，いずれも他の社会ほど大きな人口比率を獲得することない．代わりにこれらの戦略が共存もしくは循環する一方で，その協力率は図 3 (b) にあるように 0.680 に安定する．

周期協力社会以外の社会では，ダイナミクスが収束するにつれて，ある戦略が単独で最大多数を占めるようになる．これに対して周期協力社会では，いくつかの戦略の人口比率が振動し，ある一定の比率に収束する，もしくは最大多数戦略が周期的に入れ替わるサイクルが続くようになる．この戦略の人口比率における協力率はほぼ一定で，ALLC, GRIM, TFT, 1MP, OSP の 5 つの戦略を含む．ただし，g が小さいときには ALLC が，大きいときには OSP の比率が増加する傾向がある．実際，裏切りによる利得の増分が増えると ALLC のような相手を処罰しない戦略は簡単に搾取されてしまう．こうした協力の移り変わりを理解するために，そのダイナミクスをいくつか吟味する．

図 4 (a) に $g = 0.50$ および $l = 0.40$ における戦略の人口比率の時間変化を示す．それぞれの戦略はまず振動を繰り返すが，徐々に振幅は小さくなり，1000 期以降はほとんど変化しない．このとき，先に述べた 5 つの戦略が生き残っており，比率の大きい順に GRIM (0.565), TFT (0.264), ALLC (0.083), 1MP (0.039), OSP (0.035) となっている．括弧内にその戦略の人口比率を示す．また，l をわずかに大きくしても同様のダイナミクスを観察するが，それ以上大きくすると不寛容社会への移行し，GRIM がすぐにすべての人口を獲得するようになる．

図 4 (b) では，l を 0.4 に固定したまま，g を 0.8 に増加させた．このとき，5 つの戦略の人口比率は循環し，一定の比率に安定しない．その最大多数戦略はほとんどが GRIM だが，定期的に TFT の比率が GRIM より高くなる．いつ計算を打ち切るかによってどの戦略が最大多数となるかが変わるため，図 3 (a) で TFT が最大多数となる領域が飛び地を形成する．ただし，図 4 (d) の $g = 1.5, l = 0.75$ のダイナミクスが示すように，TFT が最大多数となる人口比率に収束する場合も存在する．さらに l を 0.4 に固定したまま，g を 1.2 に増加させたのが図 4 (c) である．ここでは，図 4 (b) で観察した循環はなくなり，図 4 (a) と同じように戦略の人口比率がほぼ一定に収束する．ただし，その比率は大きい順に GRIM (0.477), TFT (0.258), OSP (0.197), 1MP (0.036), ALLC (0.016) となる．図 4 (a) での ALLC の代わりに OSP が GRIM，TFT に続く人口比率を獲得する．

見間違えのない完全観測では，非協力と不寛容の 2 種類の社会しか発生しない．つまり 1MP や TFT といった非自明な協力的戦略は GRIM や ALLD といった戦略ほど利得を得られない．たとえば，1MP は ALLD のような裏切りを選ぶ頻度の高い戦略からは「自動的に」搾取される．1MP は，ALLD と対戦するとき，2 回に 1 回は協力を選んでしまい利得を下げてしまうため，GRIM より高い利得を得ることなく，人口が固定してしまう．一方で，TFT は，一度裏切られたら相手が再び協力してくるまで協力に戻ることはないため，1MP のように ALLD に搾取されることはない．しかし，26 個の戦略それぞれと対戦するとき，GRIM の方が TFT より高い利得を得る傾向にある．これ

は GRIM はいったん協力が崩れると二度と協力に戻らないことで，協力に戻ろうとする戦略から搾取することができる．一方で，TFT はいったん協力が崩れるとお互いに協力と裏切りを繰り返し始めるが，この報復の連鎖からタイミングより抜け出せるような戦略はほとんどない．この結果，1MP と同様に高い利得を達成できず，完全観測において GRIM と ALLD しか最大多数にならなくなる．

一方，ほぼ完全観測では，見間違えが起きるため，不寛容な戦略では相手を処罰しすぎてしまい，利得や協力率を下げてしまう．このため，g および l が十分小さいときは 1MP が，g が l に比べて大きいときは TFT が最大多数となりえる．1MP が最大多数となるのは，1MP が均衡を構成し，裏切られたときの損失の影響が低く抑えられるときである．ほぼ完全観測で 1MP が均衡となるのは，3 章で述べたとおり $g < 0.75$ である．1MP が最大多数となる g の上限と一致する．さらに，g が小さくなれば，1MP から逸脱したときの利得の増加分が減少するので，さらに生き残りやすくなる．一方で，1MP は ALLD をはじめとする裏切りを選びやすい戦略から搾取されやすい．たとえば ALLD と対戦すると 2 回に 1 回は裏切られて，$-l$ の利得を受け取ることになる．このため，l が大きいと集団に対して高い利得を達成しにくくなる．

TFT が最大多数となるのは，GRIM と ALLC（または OSP）と共存して周期を構成するときである．この現象はある程度 g が大きく，l が g よりも小さいときにおきる．この利得パラメータ領域では ALLD や GRIM 以外の戦略は均衡にならない．しかし，TFT はその最適反応である ALLC やそれに近い OSP との混合戦略として，GRIM と共存するようになる．これは損失 l が小さければ，TFT が全戦略の中でそこそこの利得を得るためである．実際，TFT は見間違えであろうがなかろうが相手の裏切りを見た後，すぐに裏切り返し，相手の協力から損失を取り戻さない限り協力に戻らない．このため，協力と裏切りを繰り返す報復の連鎖に陥っても，全戦略の中でそこそこの利得を得る．この結果，TFT の最適反応である ALLC やそれに近い OSP との混合戦略として，GRIM と共存するようになる．

このように，不完全私的観測下では 1MP や TFT が持つ見間違えの後に協力を回復させる仕組みが機能するようになる．したがって，わずかでもお互いに相手の行動を見間違えるというノイズが，人が協力をどのように維持するかに多様性を与えているといえる．

5. 議論

5.1 周期協力社会における三すくみ

周期協力社会では，GRIM もしくは TFT が最大多数戦略となり，OSP, 1MP, ALLC を加えた 5 つの戦略が共存する．ただ，GRIM と TFT 以外の戦略の人口比率はかなり小さい．実際，1MP の比率が 1%を超えることはめったにない．また，g が小さいときは，GRIM と TFT に次いで ALLC の人口比率が大きくなり，g が大きいときは，ALLC の代わりに OSP の人口比率が大きくなる．そこで本章では GRIM vs. ALLC vs. TFT と GRIM vs. OSP vs. TFT の 2 つの 3 戦略の組の関係を分析する．

図 5 (a) に Sigmund [14] でよく知られている ALLD vs. ALLC vs. TFT のダイナミクスを示す．図中の矢印はダイナミクスが進む方向を，黒点は安定な不動点を，白点は不安定な不動点を表す．ここではプレイヤが行動を取り間違える (trembling-hand) とき，これらの 3 戦略が三すくみを形成する．TFT の人口が少ないときは ALLD へ収束する一方で，TFT の人口が一定数を超えると 3 戦略による周期が発生する．

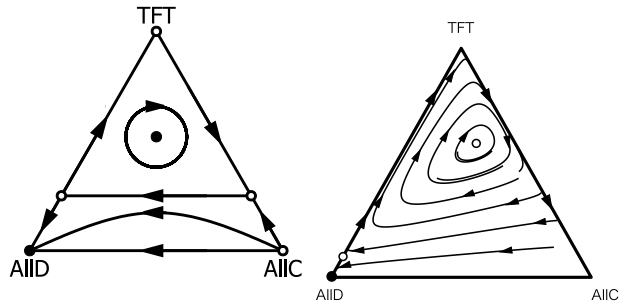
ほぼ完全観測（$p = 0.95$ および $q = 0.01$）において $g = 1.20$, $l = 0.40$ とする．図 5 (b) に，ALLD, ALLC, TFT の 3 戦略間のダイナミクスを示す．ここで不安定な不動点 (ALLD, ALLC, TFT)=(0.150, 0.266, 0.585) を中心にサイクルが形成される．図 5 (a) と同じように TFT の人口比率が一定数を下回ると ALLD に収束するようになる．次に，ALLD を GRIM に入れ替えた結果を図 5 (c) に示す．サイクルの中心が不安定な不動点 (GRIM, ALLC, TFT)=(0.527, 0.125, 0.347) に移った以外は，図 5 (b) と同じような結果となった．さらに，ALLC を OSP に入れ替えた結果を図 5 (d) に示す．ここでサイクルの中心は (GRIM, OSP, TFT)=(0.348, 0.427, 0.225) に移り，図 5 (a) とほぼ同じダイナミクスを異なる戦略の組で形成する．

すでに見てきたように周期協力社会では GRIM もしくは TFT が最大多数となる．見間違えのあるとき，TFT は ALLD や GRIM を支配する．ALLC もしくは OSP は TFT を支配するが，どちらの戦略も GRIM に支配される．26 戦略間のダイナミクスの帰結で ALLD は生き残らないが，代わりに GRIM が（行動をとり間違えるときの）ALLD の役割を果たす．また g が大きいと裏切りへの誘因が強くなるので，ALLC が生き残りにくくなるとともに，相手を 1 回だけ処罰する OSP が ALLC にとって代わる．g が大きすぎもせず，小さすぎもしないときに図 4 (b) のような複雑な周期を観測し，戦略の安定的な共存が成立しなくなる．

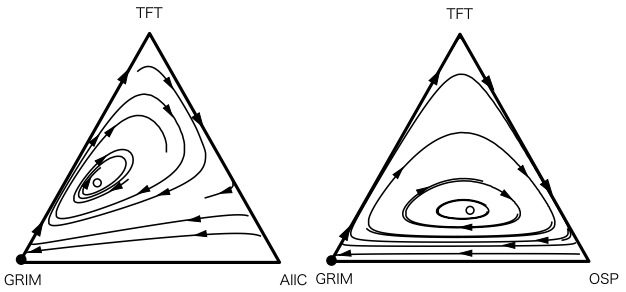
摂動とほぼ完全観測では戦略の定義が異なるため，単純に結果を比較することはできないが，摂動における ALLD vs. ALLC vs. TFT に相当する GRIM vs. OSP vs. TFT という三すくみ状態を，私的観測において世界で初めて発見した．

5.2 感度分析

本節では，利得，シグナル分布，そして突然変異率といったパラメータに関する感度分析を行う．まず，利得パラメータ g, l が期待利得に与える影響を**図 6** に示す．た



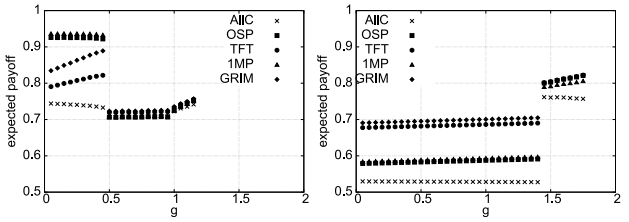
(a) ALLD vs. TFT vs. ALLC [14]　(b) ALLD vs. TFT vs. ALLC



(c) GRIM vs. TFT vs. ALLC　(d) GRIM vs. TFT vs. OSP

図 5 3 戦略間のレプリケータダイナミクス：図 5 (a) は文献 [14] より引用

Fig. 5 Replicator dynamics between three strategies: Figure 5 (a) is referred from Ref. [14].



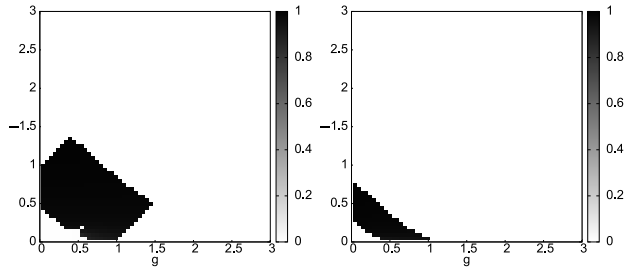
(a) l=0.20　(b) l=0.80

図 6 ゲイン g に対する収束時の期待利得

Fig. 6 Expected payoffs at convergence against gain g.

だし，ここでの期待利得は収束時の人口比における各戦略の期待利得の平均と定義する．図 6 (a) に $l = 0.2$ に固定して g を変化させたときの平均利得を示す．g が約 0.5 よりも小さい，つまり 1MP が最大多数となるとき，1MP の利得は最も高く．その値は 0.9 を超える．g が 0.5 よりも大きくなると，GRIM が最大多数を占める不寛容社会が実現する．このとき 5 戦略すべての利得が 0.7 程度になり，その差がほとんどなくなる．次に，図 6 (b) に $l = 0.8$ に固定して g を変化させたときの平均利得を示す．g が 1.4 よりも小さいとき，GRIM と TFT の利得はともに 0.7 程度であるが，GRIM が最大多数となる．g が 1.4 を超えると，GRIM, TFT, OSP の期待利得は 0.8 程度となり，1MP が 0.75 程度と若干低くなるこのときは，これら 3 つの戦略が多数を占める共存が発生する．

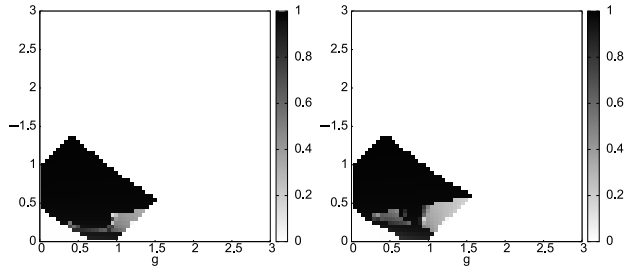
次に，シグナル分布パラメータを変化させ，ダイナミクスの帰結を観察したところ，最大多数となる戦略それぞれ



(a) p=0.90　(b) p=0.80

図 7 見間違えが起こらない確率 p に対する GRIM の比率の変化 ($q = 0.01$)

Fig. 7 Abundances of GRIM in no-error probability p when $q = 0.01$.



(a) q=0.02　(b) q=0.03

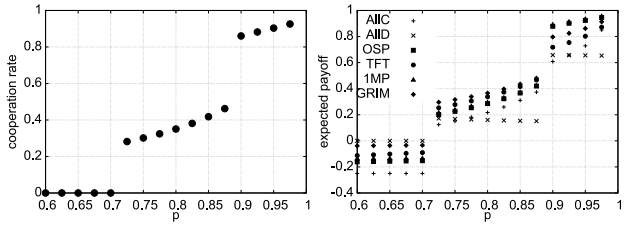
図 8 1 人が見間違える確率 q に対する GRIM の比率の変化 ($p = 0.90$)

Fig. 8 Abundances of GRIM in one-error probability q when $q = 0.90$.

の領域の大きさや，GRIM がどんな戦略と共存するかが変化した．一方で，有効な戦略の傾向に大きな変化は見られなかった．これを確認するために，**図 7** に q を 0.01 に固定したまま p を変化させたときの，**図 8** に p を 0.90 に固定したまま q を変化させたときの収束時 GRIM の人口割合を示した．ただし，GRIM が人口を獲得していない g および l が小さい領域では 1MP が，g および l 大きい領域では ALLD がほとんどすべての人口を獲得し最大多数となっている．

図 7 では p が小さくなるにつれて，ALLD が最大多数となる g および l の領域が増加していることが分かる．同じように GRIM も g および l の値が小さいとき最大多数となるが，ALLD と比べるとその領域は小さくなっている．一方で，1MP は g および l がさらに小さくないと最大多数にならなくなる．この傾向は $p = 0.95$, $q = 0.01$ とした図 3 (f) でも確認できる．そして GRIM の人口割合に注目すると，$p = 0.95$ では $g > l$ の範囲において GRIM が他戦略と共存する傾向が見られたが，$p = 0.90, 0.80$ においてはその傾向は見られなかった．

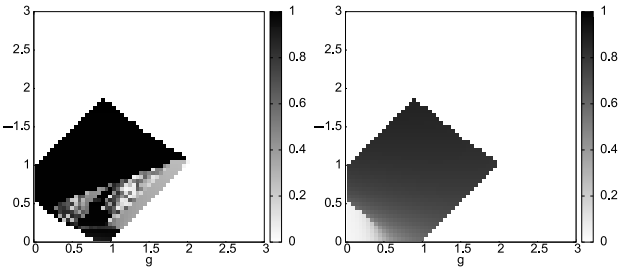
一方で，図 8 では各戦略が最大多数となる領域はほとんど変化していないが，q の値が大きくなるにつれて $g > l$ における GRIM と他戦略の共存領域が増加している．この



(a) 協力率　(b) 期待利得

図 9 見間違えが起こらない確率 p に対する協力率および利得（$g = 0.25$, $l = 0.25$）

Fig. 9 Cooperation rates and expected payoffs in no-error probability p when $g = 0.25$ and $l = 0.25$.



(a) μ=0.001　(b) μ=0.1

図 10 突然変異確率 μ に対する GRIM の比率の変化（$p = 0.95$, $q = 0.01$）

Fig. 10 Abundances of GRIM in mutation rate μ when $p = 0.95$ and $q = 0.01$.

傾向は $p = 0.90$, $q = 0.01$ とした図 7 (a) でも確認できる．

ここで再び，GRIM が他の戦略と共存しうる 3 つのパラメータ（図 3 (f), 8 (a), 8 (b)）に注目すると，両方のプレイヤが間違ったシグナルを受けとる確率 r（$r = 1-p-2q$）が比較的小さいという共通点が見つかる．具体的には $r \leq 0.06$ では GRIM が他戦略と共存するが，$r \geq 0.08$ では他戦略とほとんど共存しない．したがって，r の大きさが GRIM と他戦略の共存度合に影響を及ぼすと考えられる．

次に p が領域の変化に与える影響を見るために，$g = 0.25$, $l = 0.25$, $q = 0.01$, $\mu = 0.01$ に対して p を変化させたときの協力率と期待利得の推移を**図 9** に示す．p が大きくなる，つまりシグナルの正確さが増加するにつれて最大多数戦略が ALLD, GRIM, 1MP と変化し，協力率と期待利得が段階的に上昇する．

最後に突然変異確率の影響を吟味するため $p = 0.95$, $q = 0.01$ のほぼ完全観測で GRIM が最大多数となる領域を**図 10** に示す．突然変異確率 $\mu = 0.01$ の図 3 (f) に対して，突然変異確率を 0.001 に小さくすると，図 10 (a) にあるように $g > l$ において GRIM が他戦略と共存する領域が増加する．逆に，突然変異確率を 0.1 に大きくすると，その人口比率は安定的に大きくなる．その結果，周期協力社会で見たような最大多数戦略の周期的な変化が起こりにくくなる．いい換えると突然変異率が高いとき，多数を占めていない戦略が発生する確率が高くなる．その結果，GRIM のような不寛容な戦略の方が安定した利得を実現しやすくなるためである．

6. おわりに

本論文が対象とした私的観測下の繰り返し囚人のジレンマにおいて，ある戦略の均衡条件を解析的に求めることは現状ではほぼ不可能であることが知られている [7]．そこで，戦略空間を限定した突然変異付きレプリケータダイナミクスを用いて，私的観測下でどんな戦略が生き残るかを明らかにした．実際，完全観測では ALLD もしくは GRIM のいずれかの戦略しか生き残らなかったのに対して，不完全私的観測では見間違えの後でも協力に戻りやすい 1MP という比較的新しい戦略が生き残るようになった．さらに非自明な均衡戦略を持たないような利得構造では，TFT が GRIM, OSP とともに三すくみになり，周期協力社会を構成することを世界で始めて明らかにした．このようにして明らかにした私的観測下における多様な振舞いは今後のゲーム理論の進展に寄与することが予想される．

謝辞 本論文の執筆にあたり，神取道宏, Christian Hilbe, Martin Nowak の 3 氏から多数の有益なご助言をいただきました．また，2 人の査読者からは非専門家の視点からこちらが気が付きにくい点を多数ご指摘いただきました．ここに深く感謝いたします．本研究は JSPS 科研費 21H04890, 20K20752, 16KK0003 の助成を受けたものです．

参考文献

[1] Axelrod, R.: The Evolution of Strategies in the Iterated Prisoner's Dilemma, *Genetic Algorithms and Simulated Annealing*, Davis, L. (Ed.), pp.32–41, Morgan Kaufman (1987).
[2] Dormand, R. and Prince, P.J.: A family of embedded Runge-Kutta formulae, *Journal of Computational and Applied Mathematics*, Vol.6, No.1, pp.19–26 (1980).
[3] Hofbauer, J. and Sigmund, K.: *Evolutionary Games and Population Dynamics*, Cambridge University Press (1998).
[4] Imhof, L.A., Fudenberg, D. and Nowak, M.A.: Evolutionary cycles of cooperation and defection, *Proc. National Academy of Sciences*, Vol.102, No.31, pp.10797–10800 (2005).
[5] Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python (2001–).
[6] Kandori, M.: Repeated games, *Game theory*, Durlauf, S.N. and Blume, L.E. (Eds.), pp.286–299, Palgrave Macmillan (2010).
[7] Kandori, M. and Obara, I.: Towards a Belief-Based Theory of Repeated Games with Private Monitoring: An Application of POMDP (2010).
[8] Kraines, D. and Kraines, V.: Pavlov and the prisoner's dilemma, *Theory and Decision*, Vol.26, pp.47–79 (1989).
[9] Mailath, G. and Samuelson, L.: *Repeated Games and Reputation*, Oxford University Press (2006).
[10] Mathieu, P. and Delahaye, J.-P.: New Winning Strategies for the Iterated Prisoner's Dilemma, *Proc. 2015*

International Conference on Autonomous Agents and Multiagent Systems, AAMAS '15, pp.1665–1666 (2015).
[11] Nowak, M.: *Evolutionary Dynamics: Exploring the Equations of Life*, Harvard University Press (2006).
[12] Nowak, M. and Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit for tat in prisoner's dilemma, *Nature*, Vol.364, pp.56–58 (1993).
[13] Shampine, L.W.: Some Practical Runge-Kutta Formulas, *Mathematics of Computation*, Vol.46, No.2, pp.135–150 (1986).
[14] Sigmund, K.: *The Calculus of Selfishness*, Princeton University Press (2010).
[15] Taylor, P.D. and Jonker, L.B.: Evolutionarily Stable Strategies and Game Dynamics, *Mathematical Biosciences*, pp.145–156 (1978).
[16] Zagorsky, B.M., Reiter, J.G., Chatterjee, K. and Nowak, M.A.: Forgiver Triumphs in Alternating Prisoner's Dilemma, *PLOS ONE*, pp.1–8 (2013).
[17] 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).
[18] 関口 格：経済セミナー増刊：ゲーム理論プラス，「協調達成のための正しいお仕置きの仕方」，日本評論社 (2007).
[19] 大槻 久：有限集団における進化ゲーム理論の発展，特集「多様性と進化の統計解析」，統計数理研究所，chapter 60-2, pp.251–262 (2012).
[20] 岡田 章：ゲーム理論 新版，有斐閣 (2011).
[21] 松島 斉：繰り返しゲームの新展開：私的モニタリングによる暗黙の協調，ゲーム理論の新展開，勁草書房 (2002).
[22] 神取道宏：ミクロ経済学の力，日本評論社 (2014).
[23] 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).

付　　録

A.1 FSA リスト

本章では，図 **A·1** に本研究で使用した状態数 2 以下の 26 個の非同相な FSA のリストを示す．

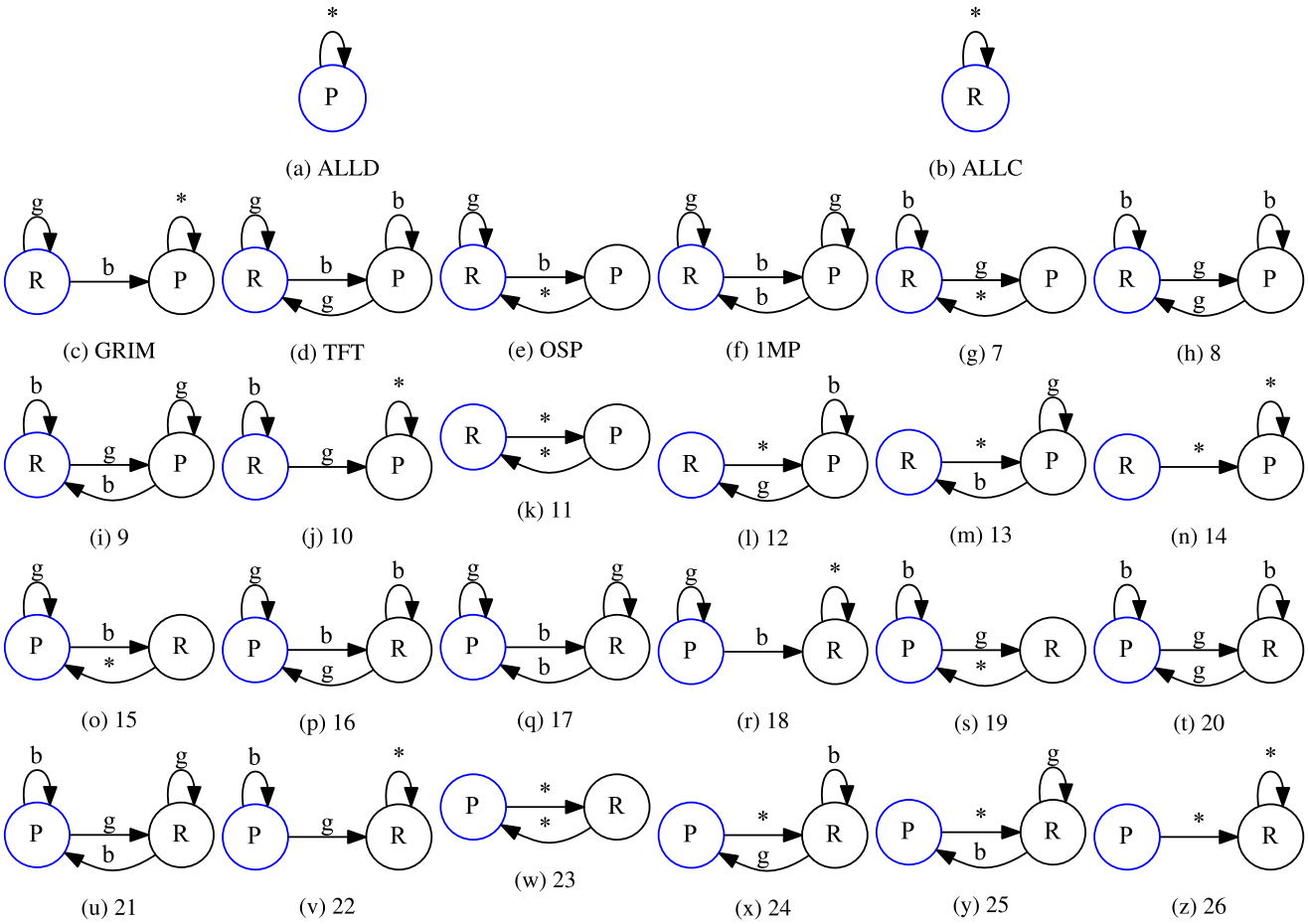


図 **A·1** 状態数 2 以下の非同相な 26 個の FSA
Fig. A·1 26 non-isomorphic FSA strategies within two states.

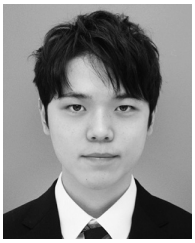
推薦文

私的観測下の繰り返し囚人のジレンマという，新規性の高い題材を対象に突然変異付きのレプリケータダイナミクスによる解析が行われている．様々な条件下におけるダイナミクスの帰結を網羅的に示し，複数の戦略が共存する場合についても詳細に説明されている．さらに，類似の条件下に囚人のジレンマゲームにおけるダイナミクスとの違いが考察されている．対象のモデルに対して詳細なシミュレーション結果と解析が行われており，新規性も高い内容であることから，推薦論文とした．

（FIT2020 第 19 回情報科学技術フォーラム
プログラム委員長　長 健太）

西野上 和真

1996 年生．2019 年 3 月電気通信大学総合情報学科卒業．2021 年 3 月同大学大学院博士前期課程修了．修士（工学）．現在，SCSK 株式会社勤務．繰り返しゲームやゲームの均衡解探索アルゴリズム等に興味を持つ．



五十嵐 瞭平　（学生会員）

1998 年生．2020 年 3 月電気通信大学情報理工学域卒業．2020 年 4 月同大学大学院博士前期課程入学．学士（工学）．繰り返しゲームやゲームの均衡解探索アルゴリズム等に興味を持つ．

岩崎 敦　（正会員）

2002 年神戸大学大学院自然科学研究科博士課程修了．同年 2004 年まで NTT コミュニケーション科学基礎研究所に勤務．2004 年九州大学大学院システム情報科学研究院助教．2013 年より電気通信大学大学院情報システム学研究科准教授．博士（学術）．ゲーム理論と最適化に関する研究に従事．オークションやマッチング等のメカニズム設計や繰り返しゲームや協力ゲーム等に興味を持つ．2020 年第 19 回情報科学技術フォーラム FIT2020 船井ベストペーパー賞，2021 年第 20 回情報科学技術フォーラム FIT2021 船井ベストペーパー賞等受賞．IEEE，人工知能学会各会員．