RA-003

相互処罰による協調: 私的観測付き無限回繰り返し囚人のジレンマの部分観測マルコフ決定過程による解法

ジョ ヨンジュン* 岩崎 敦* 神取 道宏† 小原 一郎‡
YongJoon Joe Atsushi Iwasaki Michihiro Kandori Ichiro Obara
横尾 真*
Makoto Yokoo

1 序論

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．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 [11]．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) のケースはほとんど解析されている．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) のケースが存在する．ここでは，あるプレイヤが相手の行動を間違って観測したとき，相手も間違った観測をしている，つまり，プレイヤの観測がお互いに共通するケースが集中的に研究されている．これを不完全公的観測 (imperfect public monitoring) のケースと呼ぶ [7]．近年では，これよりさらに一般的で応用範囲の広い不完全私的観測 (imperfect private monitoring) のケースへの注目が急速に高まっている [8; 14; 3].

不完全私的観測付き無限回繰り返しゲーム (infinite repeated games with private monitoring) の特徴は，あるプレイヤが相手の行動について観測したシグナルを他のプレイヤが観測できない点にある．つまり，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される．

例えば，アドホックネットワークにおける各ノードが利己的に振舞うと仮定したときのパケット転送を考える [16]．パケット転送のリクエストを受けたノードはそのパケットを転送するか（協力），破棄するか（裏切り）を選択する．もし全てのノードが協力するならば，ネットワーク全体の性能は高くなるが，他のノードが協力しているとき，自分だけ裏切ることでパケット転送にかかるコストの分，利益を増加させることができる．つまり，利己的なノードは他のノードからのリクエストを破棄する誘引をもつ．

このような状況はゲーム理論で代表的なゲームである囚人のジレンマと同じ構造をもつ．もしノードがお互いの行動を完全に観測できるなら 1 つのノードだけが裏切っても他のノードからも裏切られるので利得はあまり増加しない．しかし，現実にはノードはお互いの行動を完全には観測できない（観測にノイズが含まれる）ので，裏切るノードを的確に排除しながら，ネットワーク全体の性能を維持するような戦略が問題となる．このようにネットワーク／人工知能分野においてノイズを含む環境を扱う枠組みの重要性は増加している．実際，文献 [15; 17]では相手の行動の観測に制限が課されたエージェントによる繰り返し渋滞ゲーム (repeated congestion game) が考察されている．

相手の行動に関する観測にノイズが含まれる状況下における繰り返しゲームに関するシミュレーション研究は非常に多い [9]．しかし一方で，解析的にゲームの帰結，つまり均衡を求める研究はほとんど成果をあげられていなかった．これは，不完全私的観測付き繰り返しゲームにおける均衡を求めるには，非常に複雑な統計的推論を必要とするためである [3]．その結果，非常に制限された特定のシグナル分布しか検討できなかった．

ごく最近，均衡における振舞いを有限状態オートマトン (finite state automaton, FSA) で記述し，部分的観測可能マルコフ決定過程 (partially observable Markov decision process, POMDP) の理論を用いることで，ある FSA が均衡であるか否かを明らかにできることを文献 [4]が示した．ここで，任意のシグナル分布に対して，与えられた FSA プロファイルが均衡を構成するか否かを判定する扱いやすい計算方法を提案している．しかし，そのアルゴリズムは未解決である．

また，[4]より提案された方法を用いても，ある戦略が既存の均衡概念である部分ゲーム完全均衡を構成することを示計算量は非常に多く，その計算の停止性を保証しない．

そこで本論文では対称純粋有限状態均衡 (symmetric

* 九州大学大学院システム情報科学府情報学専攻
† 東京大学経済学部
‡ UCLA 経済学部

pure finite state equilibrium, SPFSE) という新しい均衡概念を提案し，文献 [4]の枠組みにもとづいて，この均衡を計算する具体的かつ停止性を保証するアルゴリズムを提案する．これを用いて，不完全観測下で均衡を構成する代表的な FSA である修正トリガー (grim-trigger, GT, 図 1 参照) を吟味した．[*1]本論文では GT が観測に含まれるノイズが比較的小さい領域で均衡となることを明らかにした．

GT は相手の裏切りを絶対に許さない（寛容でない）ため，その期待利得が非常に小さくなる点が問題となっている．一方で，お互いに協力を促す寛容な戦略として“しっぺ返し (tit-for-tat, TFT, 図 2)”がよく知られている．しかし，2 人のプレイヤが TFT にしたがって行動を選択する場合，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互いに協力することが極めて困難になる．つまり，(C,D) と (D,C) の状態を交互に遷移し続けることになってしまう．ここから相互協力状態に戻るには，両方のプレイヤが前の期で協力 (のシグナル) を観測するという非常に起こりにくいシグナルの組合せが生起しなければならない．

本論文では，この問題を解決するために k-期相互処罰 (k-period mutual punishment, k-MP) という新しい FSA を提案する．図 4 に 2-MP の FSA を示す．これに従うプレイヤは最初協力し，相手の裏切りを観測するとプレイヤも裏切るが, k 回連続して互いに裏切った後，相互協力状態に戻る．パラメータ k を変えることで k-MP の寛容の度合いを調整できる．

k-MP は GT と Pavlov [5]というよく知られた戦略を特殊ケースとして含む ($k=\infty$ と $k=1$)．一見，相手が裏切ったという (bad) シグナルを観測すると協力状態に戻るという k-MP の振舞いは直観に反するように見える．しかし実際には，相互処罰（互いに裏切りあう行為）を一定の期間導入することで観測にノイズが含まれていても協調を維持できるようになり，2-MP が十分広いシグナルパラメータの範囲で均衡を構成することを示した．これは GT よりは狭い範囲ではあるが，その利得は GT よりはるかに高い．さらに，状態数 3 以下の FSA に対して全探索を行い，十分に広いシグナルパラメータの範囲で均衡を構成し，かつ GT より高い利得を達成する FSA が他に存在しないことを確認した．

従来，POMDP は単一のエージェントのプランニングにおいてよく用いられている一方で，ゲーム理論は複数のエージェントの相互作用を分析するために広く用いられている．しかし，これらを相互に利用した研究はほとんどなされてなかった．実際，著名な人工知能の教科書でも，“...ゲーム理論による戦略と POMDP が計算する戦略を組み合わせるためのよい方法は未だ存在しない”と言われている [13]．そのような中で数少ない例外は文献 [1]である．彼らは本論文とは異なるクラスの均衡を用いて，その計算量を吟味している．彼らのモデルにおけるプレイヤは FSA より複雑な方法で行動を選択できるが，逆に彼らの均衡計算は非常に複雑で，それを用いて均衡を実際に計算することはほとんど現実的ではない．本論文がゲーム理論と POMDP という 2 つの独立した分野をつなぎ，これらにまたがる新しい学際的研究を活性化する 1 つの重要な契機になると考えている．

2 私的観測付き無限回繰り返しゲーム

本章では文献 [4]にもとづいて，2 人対象ゲーム (プレイヤの識別子を入れ替えても意味が変わらないゲーム) における私的観測付き無限回繰り返しゲームをモデル化する．ただし，本論文で扱う手法は n プレイヤ, 非対象ゲームに容易に拡張できる．

無限回繰り返しゲームでは, プレイヤ $i \in \{1,2\}$ は同じ成分ゲーム (stage game) を無限期間 $t=1,2,\ldots$ 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，その行動プロファイルを $\boldsymbol{a}=(a_1,a_2)\in A^2$ とする．その期におけるプレイヤ i の利得を成分ゲームの利得関数 $g_i(\boldsymbol{a})$ で与える．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．$\boldsymbol{\omega}$ を観測プロファイル $(\omega_1,\omega_2)\in\Omega^2$ とする．また, プレイヤが行動プロファイル $\boldsymbol{a}$ を選択した場合において生起するシグナルプロファイルが $\boldsymbol{\omega}$ である $o(\boldsymbol{\omega}\mid\boldsymbol{a})$ を同時確率とする．このとき，有限集合 Ω に対する $o_i(\omega_i\mid\boldsymbol{a})$ を Ω_i 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．つまり，どの行動プロファイル $\boldsymbol{a}$ に対しても，それぞれのシグナルプロファイル $\boldsymbol{\omega}$ が生起する確率は正となる．

プレイヤ i の行動 a_i とシグナル ω_i から**認識可能な利得** $\pi_i(a_i,\omega_i)$ を決定する．このため, プレイヤ i の認識利得は $g_i(\boldsymbol{a})=\sum_{\boldsymbol{\omega}\in\Omega^2}\pi_i(a_i,\omega_i)o(\boldsymbol{\omega}\mid\boldsymbol{a})$ で与えられる．この定義は認識利得 π_i が a_i と ω_i 以外の情報を含まないことを保証しており，期待利得が $\boldsymbol{a}$ のみから決定される一方で，認識利得は a_i と $\boldsymbol{\omega}$ より決定される．

以上の私的観測付き無限回繰り返しゲームのモデルは次のような小売店同士の競争を想定している．つまり，競合している 2 つの小売店をプレイヤとし，それぞれの店にある商品の価格を決める行為を行動とする．このとき，ある店の来客数をその店が観測するシグナルとすれば，このシグナルは相手の小売店が決めた価格（相手の行動）の影響を受ける．この結果，自分の店の価格と来客数がそのプレイヤの行動とシグナルとなり，認識可能な利得を決定する．

最後に，成分ゲームは無限の期間上で繰り返し行わ

[*1] g もしくは b は相手の行動 C もしくは D に関するノイズを含む観測の私的シグナル ($good$ もしくは bad) を表す．

れるので，行動プロファイル $\boldsymbol{a}^1, \boldsymbol{a}^2, \ldots$ より与えられるプレイヤ i の割引利得 G_i は割引率 $\delta \in (0,1)$ により $\sum_{t=1}^{\infty} \delta^t g_i(\boldsymbol{a}^t)$ となる．また, 割り引かれた**平均利得** (毎期の利得) を $(1-\delta)G_i$ と定義する．

2.1 繰り返しゲームの戦略と有限状態オートマトン

本節では繰り返しゲームの戦略を定義し，その戦略を有限状態オートマトン (finite state automaton, FSA) で表現する場合の均衡概念について概説する．あるプレイヤ i の t 期までの私的履歴をそのプレイヤ i の過去の行動とシグナルの記録で表し，$h_i^t = (a_i^0, \omega_i^0, \ldots, a_i^t, \omega_i^t) \in H_i^t := (A \times \Omega)^{t+1}$ とする．各プレイヤの初期行動 $\boldsymbol{a}$ を決定するためのダミー履歴として h_i^0 を導入する．ここで h_i^0 は単一集合 $\{h_i^0\}$ とする．次に，プレイヤ i の純粋戦略 s_i を，あらゆる履歴をある行動に対応させる関数として定義する．厳密には，ありうる履歴の集合 $H_i = \bigcup_{t \geq 0} H_i^t$ に関して，$s_i : H_i \to A$ とする．

FSA は繰り返しゲームにおけるプレイヤの振舞いを簡略に表現する方法として知られている．本論文では，ある FSA M を状態の集合 Θ, 初期状態 $\hat{\theta} \in \Theta$, 各状態で選択される行動 $f : \Theta \to A$, 決定的状態遷移 $T : \Theta \times \Omega \to \Theta$ に対して，$\langle \Theta, \hat{\theta}, f, T \rangle$ と定義する．ここで決定的状態遷移 $T(\theta^t, \omega^t)$ は現在の状態 θ^t および私的シグナル ω^t に対して，次の期の状態 θ^{t+1} を返す関数とする．また本論文では, 初期状態を規定しない FSA を $m = \langle \Theta, f, T \rangle$ と定義し，有限状態プレオートマトン (finite state preautomaton, pre-FSA) と呼ぶ．以上より，**対称純粋有限状態均衡** (symmetric pure finite state equilibrium, SPFSE) を定義する．

定義 1. 対称純粋有限状態均衡 *(SPFSE)* とは，各プレイヤの均衡経路上の振舞いをある *FSA* $M = \langle \Theta, \hat{\theta}, f, T \rangle$ で与えられる場合の私的観測付き繰り返しゲームの純粋戦略逐次均衡である．

ここで，逐次均衡とはナッシュ均衡の不完全情報ゲームにおける精緻化の 1 つであり，本論文で繰り返しゲームの均衡を議論するためには，ある FSA が定義するのは均衡経路上の振舞いのみでよい．この点については後述するが，詳細は文献 [4]を参照されたい．

この概念で重要なのは，ある FSA M が均衡を**構成する**ことが意味するのは，プレイヤ 2 が M にしたがって振る舞う限り，プレイヤ 1 の最適反応がその M にしたがって振舞うことになる点である．このため，本論文ではプレイヤ 1 の戦略空間を一切制限していない．つまり，FSA で表現可能な戦略だけではなく（無限の状態数を要する）全ての可能な戦略を考慮した上で，SPFSE を構成する M が最適反応となる．

2.2 POMDP としての私的観測の問題

本節では，ある FSA $M = \langle \Theta, \hat{\theta}, f, T \rangle$ が SPFSE を構成するかどうかを確認する手順を述べる．まず，各プレイヤが FSA M にしたがって行動すると仮定し，2 つの FSA の積をとると，両プレイヤの行動の対を状態とした積 FSA を作ることができる．これを用い, プレイヤ 1 の期待割引利得 $V_{\hat{\theta},\hat{\theta}}$ を以下の線形連立方程式を V_{θ_1,θ_2} に関して解くことで計算できる．

$$V_{\theta_1,\theta_2} = g_1((f(\theta_1), f(\theta_2))) + \delta \sum_{(\omega_1,\omega_2) \in \Omega^2} o((\omega_1,\omega_2) \mid (f(\theta_1), f(\theta_2))) \cdot V_{T(\theta_1,\omega_1),T(\theta_2,\omega_2)}$$

次に，プレイヤ 2 が M にしたがって行動するとき，プレイヤ 1 の最適反応をどのようにして求めるかを述べる．プレイヤ 2 がその FSA のどの状態にいるかによって表されるマルコフ過程をプレイヤ 1 は解くことになる．しかし，プレイヤ 1 はプレイヤ 2 の状態を直接観測できないため，プレイヤ 1 の最適反応を求める問題は POMDP における最適ポリシー を求める問題と等価となる．

この問題の POMDP はプレイヤ 2 の状態集合 Θ, プレイヤ 1 の行動集合 A, プレイヤ 1 の観測集合 Ω, 観測確率関数 O, 状態遷移関数 P, 利得関数 R に関して, $\langle \Theta, A, \Omega, O, P, R \rangle$ と定義される．ここで Θ, A, Ω の定義はすでに述べた．$O(\omega_1 \mid a_1, \theta^t)$ は，プレイヤ 2 が状態 θ^t にいるとき，プレイヤ 1 が行動 a_1 を行った後, ω_1 を観測する条件付き確率を表す：$O(\omega_1 \mid a_1, \theta^t) = o_1(\omega_1 \mid (a_1, f(\theta^t)))$.

標準的な POMDP のモデルでは，観測確率を次の状態 θ^{t+1} によって決定するように定義する．本論文では, 私的観測付き繰り返しゲームの定式化に合わせて観測確率を現在の状態 θ^t によって決定するように変更している．しかし，本質的な違いはないので必要に応じて使い分けることができる．

$P(\theta^{t+1} \mid \theta^t, a_1)$ が表すのは，現在の状態が θ^t およびプレイヤ 1 の行動 a_1 に対して，次状態が θ^{t+1} となる条件付き確率である：

$$P(\theta^{t+1} \mid \theta^t, a_1) = \sum_{\omega_2 \in \Omega \mid T(\theta^t,\omega_2) = \theta^{t+1}} o_2(\omega_2 \mid (a_1, f(\theta^t))).$$

最後に，期待利得関数 $R : A, S \to \mathbb{R}$ を $R(a_1, \theta^t) = g_1((a_1, f(\theta^t)))$ と定義する．

ある FSA $M = \langle \Theta, \hat{\theta}, f, T \rangle$ が SPFSE を構成する否かを確認するためのアルゴリズムを以下に示す．これは [4]のアイデアを基礎にしたアルゴリズムであり，既存の POMDP ソルバーを用いて計算を実行出来る．

1. まず, 2 人のプレイヤが M にしたがって行動するときの積 FSA の線形連立方程式を解き, プレイヤ 1 の期待割引利得 $V_{\hat{\theta},\hat{\theta}}$ を求める．
2. POMDP $\langle \Theta, A, \Omega, O, P, R \rangle$ に関して，(pre-FSA として得られる) 最適ポリシー Π^* とその価値観数を求める．ここで標準的な POMDP ソルバー，例えば [2]など，を用いる．

3. プレイヤ 2 が $\hat{\theta}$ にいるかどうかに関してプレイヤ 1 がもつ信念を $b_{\hat{\theta}}$ とする．もし，$v(b_{\hat{\theta}}) = V_{\hat{\theta},\hat{\theta}}$ ならば, その FSA $M = \langle \Theta, \hat{\theta}, f, T \rangle$ は SPFSE を構成する．

より正確に述べると $v(b_{\hat{\theta}}) = V_{\hat{\theta},\hat{\theta}}$ を確実に確認するには計算誤差の問題がある．このため，求めた最適ポリシー Π^* も本当に最適か否かを確認している．よって，もし Π^* が M の pre-FSA m と完全に一致しなくても，その FSA は SPFSE を構成できる．これはプレイヤが M にしたがって行動する場合，決して到達できない信念状態が存在するためである．pre-FSA m はそのような信念状態における最適な振舞いを記述する必要はなく，Π^* も全ての可能な信念状態における最適な振舞いを記述する必要もない．

ある FSA M が SPFSE を構成するか否かを確認するために，まず相手となるプレイヤが M にしたがって行動しているときに最適ポリシー Π^* の中から最適な初期状態 θ^* を見つける必要がある．次に, Π^* の一部分，つまり θ^* から到達できる状態の集合を吟味し，この部分が M と一致するかどうかを確認する．このとき，M はそれ自身に対する最適反応となり, SPFSE を構成する．一般には，複数の最適ポリシー が存在しうるが, POMDP ソルバーは通常ただ 1 つの最適ポリシー のみを返す．したがって，もしある最適ポリシー Π^* が m を含んでいなくても, m を含む別の最適ポリシー が存在する可能性がある．それでも，m を 1 つの初期ポリシーとして用いて，M が SPFSE を構成する限りにおいて，Π^* が m を含むことを確かめることができる．

3 ノイズを含む観測のある囚人のジレンマ

本節では，提案したアルゴリズムを無限回繰り返し囚人のジレンマに適用する．ここで成分ゲームの利得を次のように与える．

	$a_2 = C$	$a_2 = D$
$a_1 = C$	$1, 1$	$-y, 1+x$
$a_1 = D$	$1+x, -y$	$0, 0$

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，$\omega_i \in \{g, b\}$ (*good*,*bad*) とする．例えば, プレイヤ 2 が C を選んだ (協力した) 時，プレイヤ 1 が正しいシグナル $\omega_i = g$ を受け取る確率は十分高いが，間違ったシグナル $\omega_i = b$ を受け取る可能性もある状況を想定する．また，片方のプレイヤのみ間違ったシグナルを受け取る確率を $q > 0$, そして両方のプレイヤが間違ったシグナルを受け取る確率を $r > 0$ と仮定する．行動プロファイルが (C, C) の時，私的シグナルの同時分布 $o(\boldsymbol{\omega} \mid \boldsymbol{a})$ を次のように与える (行動プロファイルが (D, D) の時は p と r を入れ替える).

	$w_2 = g$	$w_2 = b$
$w_1 = g$	p	q
$w_1 = b$	q	r

同じく, 行動プロファイルが (C, D) の時, 私的シグナルの同時分布を次のように与える (行動プロファイルが (D, C) の時は p と r を入れ替える)．$p + 2q + r = 1$ の

	$w_2 = g$	$w_2 = b$
$w_1 = g$	q	r
$w_1 = b$	p	q

制約の下, $p \in (1/2, 1)$ 及び $q \in (0, 1/4)$ を仮定する．また $g_i(\boldsymbol{a})$ が定数となるように $\pi_i(a_i, \omega_i)$ を決定する．このシグナル分布は相手の行動に対するノイズを含むシグナルを自然な形で表現している．以降では特に断らない限り，$x = 1$, $y = 1$, 割引因子 $\delta = 0.9$ とする．

3.1 修正トリガー

本節では完全観測における繰り返しゲームにおける代表的な均衡戦略として，修正トリガー (grim-trigger, GT, 図 1) と呼ばれる FSA を扱う．GT は最初協力し，相手の裏切りを観測するとそれ以降裏切り続ける．この FSA は R (reward, 報酬) と P (punishment, 処罰) の 2 つの状態を持っている．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. 2 人のプレイヤがともに GT にしたがって行動する場合，その積 FSA は RR, RP, PR, PP の 4 つの状態を持つ．したがって、この積 FSA に関する線形連立方程式は

$$\begin{pmatrix} V_{RR} \\ V_{RP} \\ V_{PR} \\ V_{PP} \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 2 \\ 0 \end{pmatrix} + \delta \begin{pmatrix} p & q & q & r \\ 0 & q+r & 0 & p+q \\ 0 & 0 & q+r & p+q \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} V_{RR} \\ V_{RP} \\ V_{PR} \\ V_{PP} \end{pmatrix}$$

となり，これを解くことで,

$$V_{RR} = \frac{1 - \delta r}{(1 - \delta p)(1 - \delta r - \delta q)}$$

を得る．

図 9 に GT が SPFSE を構成するシグナルパラメータの範囲を示す．x 軸は $o((g,g)|(c,c))$ や $o((b,g)|(c,d))$ のように，シグナルの正確さ p を示す． y 軸は $o((g,b)|(c,c))$ や $o((b,g)|(d,d))$ のように，片方のプレイヤのみが間違ったシグナルを受け取る確率を示す．プレイヤが観測するシグナルは p が大きい程正確になる．つまり，相手が C (協力)/ D (裏切り) を選ぶ時，プレイヤは g/b を観測しやすい．一方で，q が小さいと 2 人のプレイヤが観測するシグナルはお互いに強い相関を持つ．例えば，プレイヤ 1 が観測するシグナルが間違っていれば, プレイヤ 2 も間違ったシグナルを観測している可能性が高くなる．

GT は基本的に p が大きく q が小さい，つまり，シグナルが比較的正確で，その相関が強い領域で SPFSE を

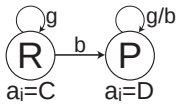

図 1　GT

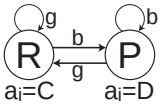

図 2　TFT

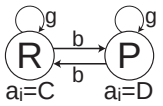

図 3　1-MP

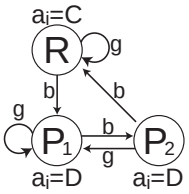

図 4　2-MP

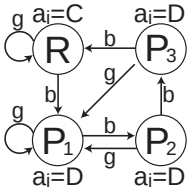

図 5　3-MP

構成する．GT が均衡を構成する領域の右側では，p および q の両方が大きくなる．つまり，シグナルは正確だが，その相関が弱くなっている．ここで，ある期においてプレイヤ 1 が b を観測すると仮定する．このとき，相手は g を観測している可能性が高いため，プレイヤ 1 はこのシグナルがほぼ確実に間違いであるとわかる．さらにこの領域ではシグナルの相関が弱いため，プレイヤ 2 は正しいシグナルを受け取りやすい．これより，プレイヤ 1 は裏切りで始めるよりは最初から協力し続ける方が良い．また，図 9 の左側の領域は，p が比較的小さいため，2 期以上ゲームが続くとプレイヤ 2 が正しいシグナルを受け取る可能性が小さくなる．したがってプレイヤ 1 は協力し続けるより裏切りで始める方が良くなる．

GT の欠点は一度でも相手からシグナル b を受け取ると 2 度と相手を許さない，つまり寛容さに欠ける点にある．例えば，$p = 0.9, q = 0.01, \delta = 0.9$ に対して，2 人のプレイヤが協力し続ける場合の期待割引利得は 10 であるのに対して, GT にしたがって行動を選ぶ場合はおよそ 5.31 にまで小さくなる．そこで，次章でより寛容な戦略の新しいクラスを提案し，吟味する．

4　k-期相互処罰

繰り返しゲームにおいて寛容な戦略として代表的なものに “しっぺ返し (tit-for-tat, TFT, 図 2)” がある．しかし，両プレイヤがしっぺ返しを取る場合，いったん相手が裏切ったというシグナルを観測すると互いに協力することが極端に困難になる．図 6 にしっぺ返しの積 FSA を示す．ここで，太線・細線・点線はそれぞれ p, q, r の確率で遷移することを意味する．本論文では p が q および r より十分大きいと仮定している．そこで p が十分大きい限り，いったん間違ったシグナルを観測すると，プレイヤは状態 (C, D) と (D, C) を繰り返すサイクルから抜け出すのが非常に難しいことを図 6 は示している．このサイクルを抜け出し (C, C) に戻るにはプレイヤは協力から逸脱する方が良い．したがって，しっぺ返しは SPFSE を構成しない．同じ理由から不完全観測下だけでなく，完全観測下でさえも部分ゲーム完全均衡を構成できない．その上，しっぺ返しの組合せが実現する利得は非常に低くなる．これは図 6 からも分かるように，いったん間違ったシグナルを観測した後，再び (C, C) に戻ることは非常に難しく，$q > 0$ かつ $r > 0$ である限り，不変分布において (C, C) が再び起こる確率は 0.25 しかないためである．

次に，しっぺ返しを少し修正し，図 3 に示す FSA を考える．本論文ではこの FSA を 1-期相互処罰 (1-MP) と呼ぶ．1-MP は，従来 *Pavlov* [5]と知られていた FSA である．この FSA の下，プレイヤは最初に協力し，相手が裏切るとプレイヤも裏切るが，互いに 1 期裏切った後，そのプレイヤは協力に戻る．図 7 に 1-MP の積 FSA を示す．片方のプレイヤのみが間違ったシグナルを観測した場合でもプレイヤはすぐに相互協力状態 RR に戻れることを確認できる．プレイヤが RR 状態にいる (不変分布に対する) 確率の期待値は $p - 2q$ となる．

しかし残念なことに，1-MP は寛容すぎるため，本論文で扱うパラメータの範囲では SPFSE を構成しない．基本的に，1-MP は相手に裏切られても 1 期だけ互いに裏切ると協力状態に戻ってしまう．このため裏切りによる利得 x が次期での損害 y と一致するため，将来の利得を割引する限り，1-MP は完全観測下でも SPFSE を構成できない．したがって，1-MP のアイデアを k-期相互処罰 (k-MP) へと一般化した．この FSA では，プレイヤは最初協力する．もし，相手が裏切ると，プレイヤも裏切る．しかし，連続して k 期互いに裏切った後，プレイヤは協力に戻る．

図 4 に 2-MP の FSA を示す．2-MP は 1-MP よりは相手の裏切りに対して厳しい（寛容ではない）が，相手が常に裏切る場合，2-MP は 3 回に 1 回は必ず協力する．k を大きくすることでこの FSA はより厳しくなり，$k = \infty$ のとき，GT と等価となる．さらに図 8 は 2-MP の積 FSA を示している．簡単のため，もっとも大きい確率 p での遷移を示す太線のみを図示している．どのようなノイズを含む観測が発生しても，プレイヤは素早く相互協力状態 RR に戻ることができる．

図 9 は 2-MP が SPFSE を構成するシグナルパラメータの範囲を示している．比較のため，GT が SPFSE を構成する範囲も示している．図 9 より $k = 2$ とするだけで，GT よりは狭いが十分広い範囲で k-MP が SPFSE を構成できることが分かる．シグナルの相関が強い場合 ($q \fallingdotseq 0$)，2-MP はシグナルが 8 割以上正確であるとき SPFSE を構成する ($p \in [0.82, 1)$)．逆にシグナルの相関が弱くなると (つまり $q > 0.04$ の範囲では), 2-MP は SPFSE を構成できなくなる．$q = 0.04$ の時, 2-MP は $p \in [0.86, 0.91)$ の範囲で SPFSE を構成する．ここで重要なのは，p が十分大きい場合，2-MP より GT の方

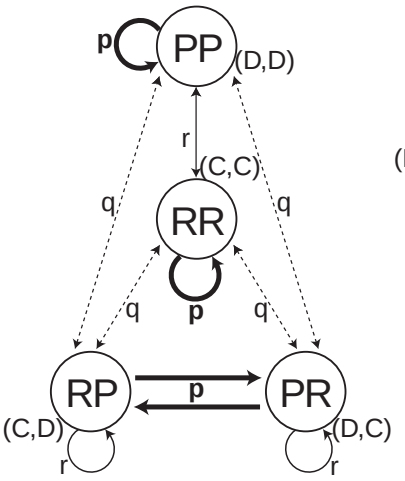

図 6　TFT の積 FSA

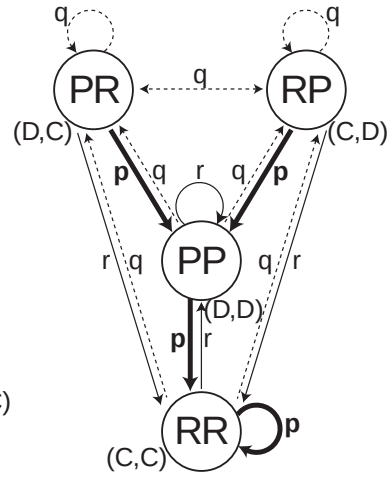

図 7　1-MP の積 FSA

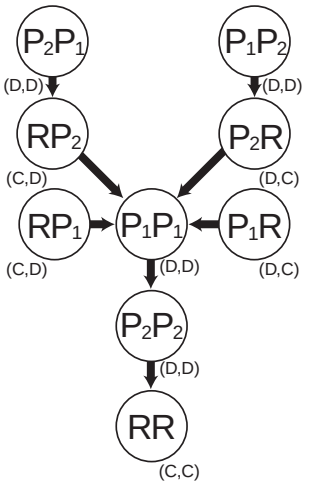

図 8　2-MP の積 FSA

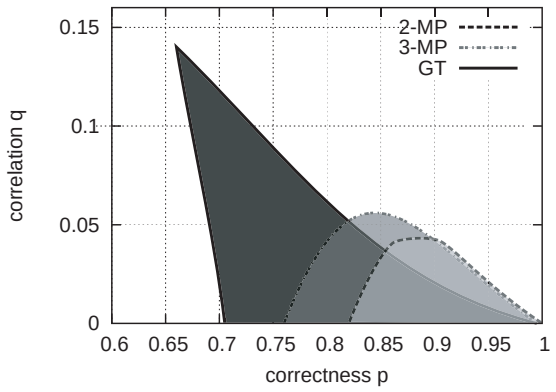

図 9　修正トリガー/2-MP/3-MP が SPFSE となるシグナル範囲. 可能なパラメータ範囲は $p+2q \leq 1$ である.

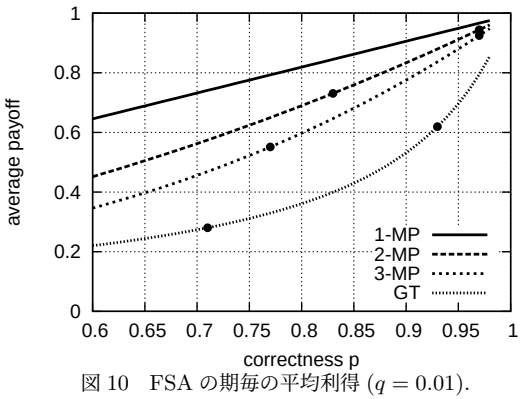

図 10　FSA の期毎の平均利得 ($q = 0.01$).

がシグナルの相関の強さの影響を受け易いことである．実際，p が 0.86 以上のとき，2-MP は SPFSE を構成する一方で GT が SPFSE を構成できない q の範囲が存在する．加えて，図 9 には 3-MP が SPFSE を構成するシグナルパラメータの範囲も示した．3-MP が SPFSE となる範囲は 2-MP より広くなっていることがわかる．

図 10 に GT と k-MP の平均利得を示す．ここでシグナルの相関 q を 0.01 に固定し，x 軸はシグナルの正確さ p，y 軸は期毎の平均利得を表す．また，平均利得が 1 になるのは相互協力が常に成立している状態を意味する．明らかに，シグナルの正確さによらず，2-MP が修正トリガー より高い平均利得を実現している．同様に，2-MP は 3-MP より高い平均利得を実現している．また，1-MP は 2-MP よりさらに高い平均利得を実現するが，本論文における利得行列の設定では SPFSE を構成できない．図のそれぞれの線にある 2 点間で，それぞれの FSA は SPFSE を構成している．ここで，k が大きくなるにつれて SPFSE を構成する p の範囲は広がるが，一方でその平均利得は低くなっている．

次に，プレイヤが将来の利得をどれだけ重要に考えるかを表す割引因子 δ の影響を示す．図 11 は割引因子 $\delta \in [0.5, 0.95]$ を変化させた時の平均利得を示している．図のそれぞれの線にある点より右側，つまり，その点がより δ が大きくなる範囲で，各 FSA は SPFSE を構成する．ここで，δ が大きくなるにつれて平均利得は減少するが，k-MP と修正トリガー が実現する平均利得の差が徐々に広がっていく．

最後に, 十分広いシグナルパラメータの範囲で SPFSE を構成でき，GT より高い平均利得を実現する FSA が k-MP 以外に存在するかどうかを吟味する．我々は状態数が 3 以下，つまり全部で $|A|^{|\Theta|} \cdot |\Theta|^{|\Theta| \cdot |\Omega|}$=5832 個の FSA を数え上げて吟味した．この結果，十分なシグナルパラメータの範囲で SPFSE を構成する FSA を 11 個発見した（ただし，実質的に同じ FSA になるものを除

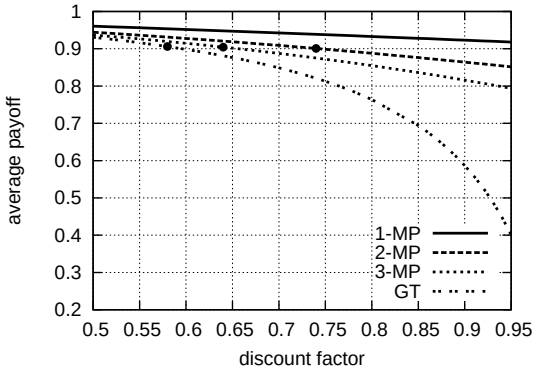

図 11　割引因子の影響 (p=0.92, q=0.01).

いている）．しかし，それらの中で GT より高い平均利得を達成する FSA は 2-MP しかないことがわかった．

5　議論

従来，1-MP は Pavlov や “win-stay, lose-shift [10]” などと呼ばれている．Pavlov は進化シミュレーションの分野でよく扱われている (例えば [9])．そこでは，私的観測とは異なるノイズ，すなわちプレイヤが選択した行動を間違うことがある場合 (*trembling hands*) の繰り返し囚人のジレンマにおける Pavlov の様々な拡張を吟味している．一方で，完全観測付き繰り返し囚人のジレンマでは，割引因子が十分高い場合，Pavlov が部分ゲーム完全ナッシュ均衡を構成することが知られているが，しかしながら，著者らが知る限り，1-MP/Pavlov は私的観測付き繰り返しゲームで均衡を構成することはこれまで証明されていない．

本論文の均衡分析は，裏切りによる利得とノイズが十分小さい場合，1-MP は 2-MP より効率的な均衡を構成しうることを示している．本論文の利得行列では，各期での裏切りによる利得は次期での裏切りによる損害と同じであるため，1-MP は SPFSE を構成できない．裏切りによる利得 x が 0.8 以下の場合，私的シグナルのパラメータのある範囲で 1-MP が SPFSE を構成できることをすでに確認している．

さらに本論文は相手の行動に関するノイズを含むシグナルを自然に記述できる観測構造を仮定した．この構造はプレイヤは (C, D) の後 (b, g) を観測しやすくなっている．一方で，私的観測下での純粋戦略均衡を解析した数少ない文献 [6; 12]では，(C, D) の後，主に (g, g) や (b, b) を受け取るような，プレイヤがほぼ同じシグナルを受け取る**ほぼ公的観測**(almost public monitoring) を扱っており，本論文で扱った観測構造とは異なる．こうした構造における均衡解析は今後の課題である．

6　結論

本論文は SPFSE という新しい均衡概念及び，それを解く具体的なアルゴリズムを提案し，これらを用いて相手の行動に関してノイズを含むシグナルをプレイヤが受け取る私的観測付き無限回繰り返し囚人のジレンマにおいてシンプルな純粋戦略均衡を明らかにした．従来，私的観測付き無限回繰り返しゲームの均衡解析は非常に難しい問題とされてきた．しかし，本論文で提案した POMDP ソルバーを用いたアルゴリズムは，均衡における振舞いを記述した FSA が SPFSE を構成するか否かを調べることを可能した．そこでまず，ノイズを含む観測が修正トリガー の振舞いに与える影響を吟味した．さらに，新しい戦略のクラスである k-MP 戦略を提案し，この戦略が修正トリガー と Pavlov [5]というよく知られた戦略をその特殊なケースとして含むことを示した．その上で，k-MP 戦略が十分広いシグナルパラメータの範囲で SPFSE を構成し，修正トリガーより高い利得を実現することを明らかにした．加えて，本論文では状態数 3 以下の FSA に対して全探索を行い，十分に広いシグナルパラメータの範囲で均衡を構成し，かつ修正トリガー より高い利得を達成する FSA が他に存在しないことを確認した．

本論文で提案したアルゴリズムは私的観測付き繰り返しゲームの均衡解析に新たな展開を与えうる．例えば，行動に関わらずプレイヤらがお互いに共通のシグナルを観測するようなほぼ公的観測なケースを従来より詳細に吟味できる．実際，このケースにおいて 2-MP は SPFSE を構成しないことを部分的に明らかにしている．また，利己的なエージェントによるパケットルーティング問題などをモデル化した渋滞ゲームにノイズのある観測を導入することで，より現実に近い状況を分析していきたい．

謝辞

本研究の遂行にあたり，日本学術振興会科学研究費補助金基盤研究 (A) (課題番号 20240015) の助成を受けました．ここに深く感謝いたします．また，非常に有益なコメントを下さった電気情報通信学会 情報・システムソサイエティアルゴリズム (AL) の 2 名の査読者に深く感謝いたします．

参考文献

[1] Prashant Doshi and Piotr J. Gmytrasiewicz. On the Difficulty of Achieving Equilibrium in Interactive POMDPs. In *Proceedings of the 21st National Conference on Artificial intelligence*, pp. 1131–1136, 2006.

[2] Leslie Pack Kaelbling, Michael L. Littman, and

Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, Vol. 101, pp. 99–134, 1998.

[3] Michihiro Kandori. *Game theory*, Repeated games, pp. 286–299. Palgrave macmillan, 2010.

[4] Michihiro Kandori and Ichiro Obara. Towards a Belief-Based Theory of Repeated Games with Private Monitoring: An Application of POMDP. mimeo, 2010.

[5] David Kraines and Vivian Kraines. Pavlov and the prisoner's dilemma. *Theory and Decision*, Vol. 26, pp. 47–79, 1989.

[6] George J. Mailath and Stephen Morris. Repeated games with almost-public monitoring. *Journal of Economic Theory*, Vol. 102, No. 1, pp. 189–228, 2002.

[7] George Mailath and Larry Samuelson. *Repeated games and reputations: long-term relationships*. Oxford University Press, 2006.

[8] 松島斉. ゲーム理論の新展開, 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, pp. 89–114. 勁草書房, 2002.

[9] Martin Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, 2006.

[10] Martin Nowak and Karl Sigmund. A strategy of win-stay, lose-shift that outperforms tit for tat in prisoner's dilemma. *Nature*, Vol. 364, pp. 56–58, 1993.

[11] 岡田章. ゲーム理論. 有斐閣, 1996.

[12] Christopher Phelan and Andrzej Skrzypacz. Beliefs and Private Monitoring. mimeo, 2009.

[13] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, 2002.

[14] 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, pp. 106–109. 日本評論社, 2007.

[15] Moshe Tennenholtz and Aviv Zohar. Learning equilibria in repeated congestion games. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, pp. 233–240, 2009.

[16] Wenjing Wang, M. Chatterjee, and K. Kwiat. Cooperation in ad hoc networks with noisy channels. In *Proceedings of the 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, pp. 1–9, 2009.

[17] 山田陽介, 小野廣隆, 来嶋秀治, 山下雅史. ある種の不完全情報渋滞ゲームの近似的ナッシュ遷移の収束性. 第 9 回情報科学技術フォーラム (FIT2010), pp. 232–233, 2010.