情報処理学会論文誌 Vol.53 No.11 2445–2456 (Nov. 2012)

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返しゲームにおける均衡分析プログラム

ジョ ヨンジュン[1,a)] 岩崎 敦[1,b)] 神取 道宏[2,c)] 小原 一郎[3,d)] 横尾 真[1,e)]

受付日 2012年1月27日, 採録日 2012年7月2日

概要：本論文では不完全私的観測付き繰返しゲームの均衡を分析するプログラムを提案する．不完全私的観測付き繰返しゲームは，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持つ．こうしたゲームは人工知能や経済の分野において様々な適用領域を持つため，大きく注目されている．しかし，この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が必要になるため，従来難しい未解決問題として知られていた．近年，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 が均衡を構成するかどうかを明らかにできることが示された．しかし，その具体的な実装方法や実際の問題へ適用するためのプログラムは提供されていない．そこで本論文ではまず，標準的な POMDP ソルバのラッパとなるプログラムを開発する．このプログラムでは私的観測付き繰返しゲームの記述と FSA を入力として，その FSA が対称的均衡を構成するかどうかを自動的に確認できる．さらに，このプログラムを繰返し囚人のジレンマに適用し，k-期相互処罰（k-MP）と呼ぶ新しい FSA のクラスを発見した．k-MP におけるプレイヤは，初めに協力し相手の裏切りを観測するとそれ以降自分も裏切るが，続けて k 回裏切りを観測すると元に戻り協力する．このプログラムを用いて状態数 3 以下の FSA を全探索した結果，繰返しゲームにおける観測構造パラメータのいくらかの範囲で，2-MP が他の純粋戦略均衡より優れており，従来よく知られている均衡である無限期罰則のトリガ戦略（grim-trigger）よりも効率的，つまり高い平均利得を実現することが分かった．

キーワード：ゲーム理論，繰返しゲーム，部分観測可能マルコフ決定過程，有限状態機械

Equilibrium Analysis Program of Repeated Games with Private Monitoring: A POMDP Approach

Yongjoon Joe[1,a)] Atsushi Iwasaki[1,b)] Michihiro Kandori[2,c)]
Ichiro Obara[3,d)] Makoto Yokoo[1,e)]

Received: January 27, 2012, Accepted: July 2, 2012

Abstract: The present paper investigates repeated games with *imperfect private monitoring*, where each player privately receives a noisy observation (signal) of the opponent's action. Such games have been paid considerable attention in the AI and economics literature. Since players do not share common information in such a game, characterizing players' optimal behavior is substantially complex. As a result, identifying pure strategy equilibria in this class has been known as a hard open problem. Recently, Kandori and Obara (2010) showed that the theory of partially observable Markov decision processes (POMDP) can be applied to identify a class of equilibria where the equilibrium behavior can be described by a finite state automaton (FSA). However, they did not provide a practical method or a program to apply their general idea to actual problems. We first develop a program that acts as a wrapper of a standard POMDP solver, which takes a description of a repeated game with private monitoring and an FSA as inputs, and automatically checks whether the FSA constitutes a symmetric equilibrium. We apply our program to repeated Prisoner's dilemma and find a novel class of FSA, which we call k-period mutual punishment (k-MP). The k-MP starts with cooperation and defects after observing a defection. It restores cooperation after observing defections k-times in a row. Our program enables us to exhaustively search for all FSAs with at most three states, and we found that 2-MP beats all the other pure strategy equilibria with at most three states for some range of parameter values and it is more efficient in an equilibrium than the grim-trigger.

Keywords: game theory, repeated games, partially observable Markov decision process, finite state automaton

1. 序論

無限回繰返しゲームは，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，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，近年注目を集めている [4], [8], [13]．不完全私的観測付き無限回繰返しゲーム（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．

たとえば，アドホックネットワークにおける各ノードが異なるプレイヤによって所有され，それぞれが利己的に振る舞うと仮定したときのパケット転送を考える [15]．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．もしすべ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という誘因を持つ．

このような状況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と同じ構造を持つ．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．たとえば，観測にノイズが含まれるため，どのノードが裏切ったかを正確に把握できないので，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 [14], [16] 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返し渋滞ゲーム（repeated congestion game）が考察されている．

このような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返しゲームにおいてプレイヤは情報を共有できない，つまり，プレイヤは自分の行動から相手が私的に観測するシグナルを観測することができないので，相手の私的シグナルについての統計的推論を必要とするためである．たとえプレイヤが比較的単純な戦略をとるとしても，その推論は途端に非常に複雑なものとなってしまう [4]．その結果，非常に制限された特定のノイズ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 が均衡を構成するかどうかを明らかにできることを文献 [5] 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どうかを判定する扱いやすい計算方法を提案している．これは，一般的な POMDP ソルバを用いて繰返しゲームの均衡が網羅的に分析できることを示唆しており，ゲーム理論と POMDP という 2 つの分野をつなぐ非常に興味深い成果である．

POMDP は単一エージェント，もしくはお互いに協調するエージェントの集団を仮定して，全体の利得の合計を最大化する行動のプランニングによく用いられる手法である．一方で，ゲーム理論は競争的な複数のエージェントの集団を仮定して，それぞれの利己的エージェントの振舞いの帰結（均衡）を求める手法である．しかし，著名な人工知能の教科書である文献 [12] でも “. . . game theory has been used primarily to analyze environments that are at equilibrium, rather than to control agents within an environment” とあるように，これらを相互に利用した研究はほとんどなかった．

数少ない例外として文献 [1], [3] があげられる．文献 [1] では，主観的均衡（subjective equilibrium）と呼ばれる均衡の計算量を吟味している．主観的均衡においてプレイヤは相手の戦略を完全には知ることができない．その結果，主観的均衡の定義は複雑になり，私的観測付き繰返しゲームにおいて主観的均衡となる戦略を現実的な時間で計算することは非常に難しいことが示されている．また，文献 [3]

[1] 九州大学大学院システム情報科学府
Graduate School of ISEE, Kyushu University, Fukuoka 819–0395, Japan
[2] 東京大学大学院経済学研究科
Faculty of Economics, The University of Tokyo, Bunkyo, Tokyo 113–0033, Japan
[3] UCLA 経済学部
Department of Economics, UCLA, Los Angeles, California 90095, USA
a) yongjoon@agent.inf.kyushu-u.ac.jp
b) iwasaki@inf.kyushu-u.ac.jp
c) kandori@e.u-tokyo.ac.jp
d) ichiro.obara@gmail.com
e) yokoo@inf.kyushu-u.ac.jp

では，部分観測可能確率過程ゲーム（partially observable stochastic games，POSGs）を扱い，被支配戦略を反復的に取り除くアルゴリズムを提案している．POSGs はエージェントは各期ごとに異なる成分ゲームに参加する点で，私的観測付き繰返しゲームの一般化であると見なせる．しかし，このアルゴリズムは有限回繰返しゲームのみに適用可能であるため，本論文で扱う無限回繰返しゲームにおける均衡を計算できない．

人工知能やマルチエージェント分野では文献 [5] の成果は残念ながらほとんど知られていない．さらに，著者らの知る限り，ミクロ経済学/ゲーム理論の分野でも，この革新的な手法を用いて繰返しゲームの均衡を構成する戦略を実際に求めた研究はない．文献 [5] には POMDP に基づく大まかな理論的概念は示されているものの，私的観測付き繰返しゲームの均衡を計算する方法を具体的に示していないことが主たる問題といえる．加えて，この手法が現実的かつ意味のある応用領域における十分複雑な例を本当に分析できるかどうかが確認されていない．とくに，POMDP モデルと私的観測付き繰返しゲームのモデルには 1 つ重要な違いが存在する．具体的には，標準的な POMDP モデルでは，ある期の観測は現在の行動（その期における行動）と次に遷移する状態（その次の期における状態）によって決まる．一方で，繰返しゲームのモデルでは，ある期の観測は現在の行動と状態（その期における状態）によって決まる．このため，文献 [5] の結果を利用・拡張することはゲーム理論の専門家にとっても，人工知能/マルチエージェントの研究者にとっても簡単ではない．

そこで，本論文ではまず，標準的な POMDP ソルバのラッパとなるプログラムを開発する．このプログラムでは私的観測付き繰返しゲームの記述と 1 つの FSA を入力とし，上で示したモデルの違いを考慮しつつ自動的に POMDP ソルバへの入力を作成する．次に，作成した入力を用いて POMDP ソルバを実行し，得られた結果と最初に入力した FSA とを比較し，その FSA が何らかの対称的均衡を構成するかどうかの答えを出力する．

さらに，このプログラムの有用性を示すため，無限回繰返し囚人のジレンマに対し，プレイヤが相手の行動についてノイズを含んだ私的シグナルを受け取るとき，どんな振舞いを記述した FSA の組が均衡を構成するかどうかを明らかにする．ある FSA の組が均衡を構成するとき，それらに従うプレイヤの利得が最大化されるため，各プレイヤはその FSA で決められた以外の振舞いに逸脱する誘因を持たなくなる．したがって，均衡を構成する FSA の組はきわめて安定な状態にあると考えられる．

本論文では，間違った観測を得る確率が小さい場合を考える．このようなシグナル分布は “ほぼ完全” 観測（nearly-perfect monitoring）と呼ばれる構造を持つ．この観測構造はきわめて自然であるにもかかわらず，本論文で開発したプログラムと POMDP ソルバを組み合わせなければ，この構造における均衡を網羅的に探索するのは不可能であった．そこで，状態数の少ない FSA について全探索した結果，k-期相互処罰（k-MP）と呼ばれる新しい FSA のクラスの発見に成功した．この FSA に従って振る舞うプレイヤは最初に協力し，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，協力に戻る．パラメータ k を変えることで k-MP の寛容さの度合いを調整できる．k-MP は “無限期罰則のトリガ戦略”（grim-trigger，GT，**図 1**）と “Pavlov” [6] 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 $k = 1$）．

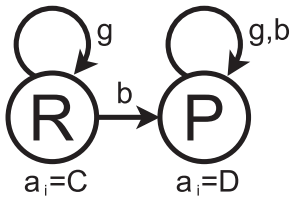

図 1 無期限罰則のトリガ戦略（GT）
Fig. 1 GT.

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，ほぼ完全観測の下で，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により，観測にノイズが含まれていても協調を維持できる．このように，本論文で提案するプログラムは異なる私的観測構造において，どのようにプレイヤはお互いの振舞いを調整するかに関する重要な知見を与えることができる．

2. 私的観測付き繰返しゲーム

2.1 モデル

本節では文献 [5] に基づいて，2 人対称ゲーム（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，非対称ゲームに容易に拡張できる．

無限回繰返しゲームでは，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．$\boldsymbol{\omega}$ をシグナルプロファイル $(\omega_1, \omega_2) \in \Omega^2$ 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 が認識できる情報である行動 a_i と観測したシグナル ω_i にだけ依存する“認識利得”（recognized payoff）$\pi_i(a_i,\omega_i)$ を決定する．プレイヤ i の認識利得は，ある特定の利得関数とシグナルの分布を事前に与え，$g_i(\boldsymbol{a}) = \sum_{\boldsymbol{\omega}\in\Omega^2} \pi_i(a_i,\omega_i)o(\boldsymbol{\omega} \mid \boldsymbol{a})$ を満たすように決定される．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．

私的観測付き無限回繰返しゲームは次のような小売店どうしの競争をモデル化している．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．

2.2 繰返しゲームの戦略と有限状態オートマトン

本節では繰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．

Definition1 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返しゲームの純粋戦略逐次均衡（pure-strategy sequential equilibrium）である．

逐次均衡とはナッシュ均衡の不完全情報動的ゲームにおける精緻化の 1 つである．従来は，均衡経路上だけでなく，均衡経路外の振舞いを記述した FSA を用いて，繰返しゲームの均衡が議論されてきた．しかし，SPFSE を導入することで，均衡経路上の振舞いのみを記述した FSA で繰返しゲームの均衡が議論できる．詳細は文献 [5] を参照されたいが，SPFSE は有限の状態数しか持たない FSA に限定した均衡概念ではない点が重要となる．つまり，ある FSA M が均衡を“構成する”とき，プレイヤ 2 が M に従って振る舞う限り，プレイヤ 1 もその M に従って振る舞うことが無限の状態数を要する FSA も考慮したうえでの最適反応になっている．

2.3 繰返し囚人のジレンマにおける観測構造

本節では，提案したアルゴリズムを無限回繰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．ここで，シグナル g は行動 C に，b は D に対応する．たとえば，プレイヤ 2 が C を選んだ（協力した）とき，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．

次に各プレイヤの私的シグナルの同時分布を $o(\boldsymbol{\omega}\mid\boldsymbol{a})$ と定義する．行動プロファイルが (C,C) のとき，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,D) のときは p と r を入れ替える）．

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

ここで，プレイヤ 1 と 2 がシグナルプロファイル (g,g) を観測する（両方とも *good* を観測する）確率を p，(g,b) を観測する確率を q とする．

同じく，行動プロファイルが (C,D) のとき，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) のときは v と u を入れ替える）．

	$w_2 = g$	$w_2 = b$
$w_1 = g$	t	u
$w_1 = b$	v	w

これらの同時分布には，$p + q + r + s = 1$ および $t + u + v + w = 1$ という制約のみが与えられる．

私的観測付き繰返しゲームは，従来の不完全観測付き無限回繰返しゲームの一般化である．シグナルパラメータを変化させることで，私的シグナルの同時分布は繰返しゲームにおけるあらゆる観測構造を表現できる．既存の観測構造には次のようなものがある．まず最初に，各プレイヤが相手の行動を完全に観測する，すなわち $p = v = 1$ であり $q = r = s = t = u = w = 0$ である場合，観測が “完全” である（perfect monitoring）と呼ぶ．次に，各プレイヤがつねに共通のシグナルを観測する，すなわち $p + s = t + w = 1$ であり $q = r = u = v = 0$ である場合，観測が “公的” である（public monitoring）と呼ぶ．

この観測構造は天気予報のニュースや会社の売上げなどのように各プレイヤが同じシグナルを観測する場合をモデル化している．公的観測には膨大な先行研究があるが，現実には同じシグナルを観測しても，そのシグナルをどうとらえるかはプレイヤごとに異なることがある．そこで，公的観測の一般化として私的観測（private monitoring）があるが，公的観測をわずかにゆるめた観測構造でしたか網羅的な分析は行われていない．具体的には，両方のプレイヤが十分高い確率で同じシグナルを観測する，（たとえば (C, D) が起こった後，プレイヤは (g, g) または (b, b) を十分高い確率で観測する）すなわち，$p + s = t + w \approx 1$ であり $q = r = u = v \approx 0$ である場合，観測が “ほぼ公的” である（almost-public monitoring）と呼ぶ．

2.4 既存の有限状態オートマトン

本節では，繰返しゲームにおいてよく知られている既存の FSA について概説する．まず最初に，有名な FSA として “無限期罰則のトリガ戦略”（grim-trigger，GT，図 1）がある．GT 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．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．多くの場合，GT は完全観測，不完全観測の両方の下で均衡を構成できる．

次に，別の有名な FSA として “しっぺ返し”（tit-for-tat，TFT，**図 2**）がある．TFT では，完全観測において，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できなくなる（したがって，TFT はサブゲーム完全ナッシュ均衡を構成しない）．

最後に，“1-期相互処罰”（1-MP，**図 3**）がある．1-MP は，従来 “Pavlov” [6] または “win-stay, lose-shift” [10] として知られている 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．

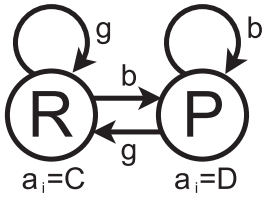

図 2 しっぺ返し（TFT）
Fig. 2 TFT.

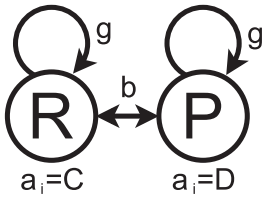

図 3 1-期相互処罰（1-MP）
Fig. 3 1-MP.

Pavlov は進化シミュレーションの分野でよく扱われている（たとえば文献 [6], [10] など）．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えることがある場合（“trembling hands”）の繰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，完全観測の下，Pavlov がサブ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．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．本論文で扱う観測構造での TFT と 1-MP についての議論を 4 章で行う．

3. 均衡分析のためのプログラム

本章では，本論文で提案するプログラム（**図 4**）について説明する．このプログラムは，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できる．

3.1 プログラムの主たる構成要素

図 4 に示す提案プログラムの主たる構成要素である “Equilibrium Analyzer” と “Standard POMDP solver” について説明する．まず，各プレイヤが FSA M に従って振る舞うと仮定し，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．

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

次に，プレイヤ 2 が M に従って振る舞うとき，プレイ

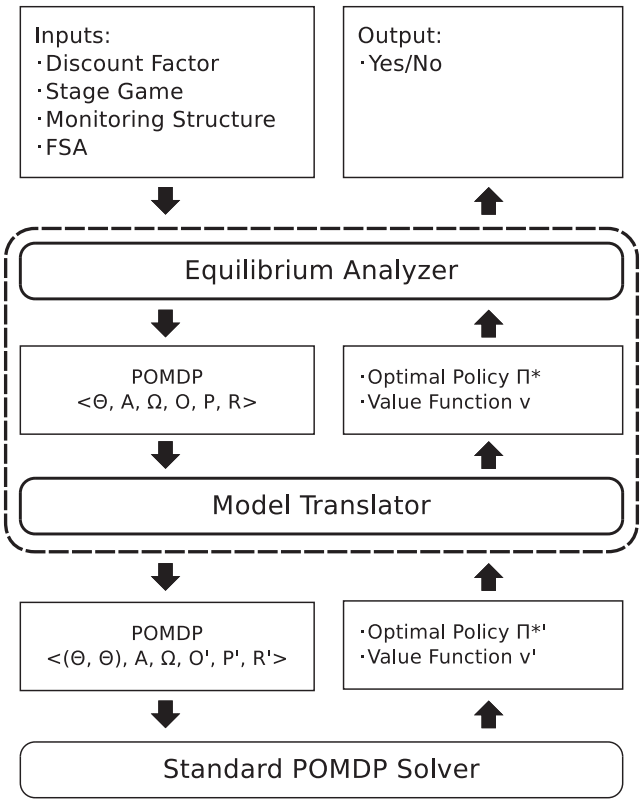

図 4 提案プログラムの流れ
Fig. 4 Flow of the program.

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 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．このような繰返しゲームにおけるモデルを標準的な POMDP モデルの定式化に変換する方法を次節で説明する．

$P(\theta^{t+1} \mid \theta^t, a_1)$ 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，次の状態が 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．これは文献 [5] のアイデアを基礎にしたアルゴリズムであり，既存の POMDP ソルバを用いて計算を実行できる．

(1) まず，2 人のプレイヤが M に従って振る舞うときの積 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 Π^* とその価値観数を求める．一般的に，この計算は収束せず最適なポリシが得られない可能性がある．このような場合，計算を終了させ準最適なポリシを pre-FSA として得る[*1]．
(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 が持つ信念を $b_{\hat{\theta}}$ 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，桁落ちの問題があるため，$v(b_{\hat{\theta}}) = V_{\hat{\theta},\hat{\theta}}$ が成立するか否かの確認が難しくなることがある．この問題を回避するために，求めた最適ポリシ Π^* と M の pre-FSA m が一致するか否かも確認する．ただし，Π^* が M の pre-FSA m と完全に一致しなくても，その FSA が SPFSE を構成することがある．これはプレイヤが M に従って振る舞う場合，到達することのない信念状態が存在しうるためである．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がないが，一方で Π^* には，到達することのない信念状態も含めて，すべての可能な信念状態における最適な振舞いが記述されている．

ある FSA M が SPFSE を構成するか否かを確認するには，まず相手となるプレイヤが M に従って行動しているときに最適ポリシ 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，複数の最適ポリシが存在しうるが，本論文で使用した POMDP ソルバは複数のポリシを結合したただ 1 つの最適ポリシのみを返す．この問題に対して，m を初期ポリシとして用いて，M が SPFSE を構成する限りは Π^* が m を含んでいることを確認する．

POMDP ソルバの計算量は一般には PSPACE 完全となるため，計算量やメモリ量はゲームのプレイヤ，手番，シグナルの数に対して指数的に増加する．こうした計算量の

[*1] 得られたポリシ（FSA）が準最適ではあるが $v(b_{\hat{\theta}}) = V_{\hat{\theta},\hat{\theta}}$ が成立しているとき，$v(b_{\hat{\theta}})$ が最適なポリシ（FSA でない）と同じであるかどうかを確認する．本手法は動的計画法に基づくため，十分な回数を繰り返し計算することで価値関数の値が最適に近づくことが保証される．文献 [5] は計算を終了させる基準として，その価値関数の上限を明らかにしている．

厳密な検証は今後の課題とするが，プレイヤ数 3，4 人程度のゲームが現時点での限界となっている．

3.2 モデルの変換

本節では “Model Translator” (図 4) について，繰返しゲームにおけるモデル $\langle \Theta, A, \Omega, O, P, R\rangle$ を標準的な POMDP モデル $\langle \Theta', A, \Omega, O', P', R'\rangle$ へ変換する方法を説明する．この 2 つのモデルでは，起こりうる行動の集合 A と観測の集合 Ω は共通している．

この変換方法の鍵となるアイデアとして複合した状態 Θ'（$\Theta' = \Theta^2$）を新しく導入する．すなわち，標準的な POMDP モデルの状態 θ'^t は，前節で示したモデルにおける 1 つ前と現在の状態 (θ^{t-1}, θ^t) の組合せを表していると仮定する．たとえば，プレイヤ 1 が GT (図 1) に従って振る舞うとき，繰返しゲームにおけるモデルでは 2 つの状態が存在する．このとき，標準的な POMDP モデルでは $2 \times 2 = 4$，すなわち，$\Theta' = \{(R,R),(R,P),(P,R),(P,P)\}$ の 4 つの状態が存在すると考える．ただし，これらの 4 つの状態において，(P,R) は実行不可能であるため考えなくてよい．

新しい状態遷移関数 $P'(\theta'^{t+1} \mid \theta'^t, a_1)$ は，$\theta'^{t+1} = (\theta^t, \theta^{t+1})$ と $\theta'^t = (\theta^{t-1}, \theta^t)$ が成立する，つまり，θ'^{t+1} における過去の状態と θ'^t における現在の状態が等しいとき，$P(\theta^{t+1} \mid \theta^t, a_1)$ と等価になり，そうでないときは 0 になる．次に，$O'(\omega_1 \mid a_1, (\theta^t, \theta^{t+1}))$ の定義について述べる．これは，状態が θ^t から θ^{t+1} に遷移するときに観測が ω_1 だったときの事後確率と等しい．したがって次のように定義できる：

$$O'(\omega_1 \mid a_1, (\theta^t, \theta^{t+1})) = \frac{\sum_{\omega_2 \in \Omega'} O(\omega_1, \omega_2 \mid (a_1, f(\theta^t)))}{\sum_{\omega \in \Omega} \sum_{\omega_2 \in \Omega'} O(\omega, \omega_2 \mid (a_1, f(\theta^t)))}.$$

ここで $\Omega' = \{\omega_2 \mid T(\theta^t, \omega_2) = \theta^{t+1}\}$ である．たとえば，GT に従って振る舞うプレイヤ 2 が状態 (R,R) にいるときにプレイヤ 1 が $a_1 = C$ を行う場合を考える．このときプレイヤ 1 が $w_1 = g$ を観測する確率は次のように与えられる

$$O'(g \mid C, (R,R)) = \frac{O(g,g \mid (C,C))}{O(g,g \mid (C,C)) + O(b,g \mid (C,C))}.$$

最後に，期待利得関数 $R'(a_1, (\theta^{t-1}, \theta^t)) = R(a_1, \theta^t)$ となる．

このモデルの変換は得られる最適ポリシに影響を与えない．つまり，変換された POMDP$\langle \Theta', A, \Omega, O', P', R'\rangle$ を解くことで，最適なポリシ Π'^*（pre-FSA で表される）とその価値関数 $v'(b_{\theta'})$ を得る．このとき，繰返しゲームにおけるモデルでの最適ポリシ Π^* は，Π'^* と必ず等しくなる．また，$\theta' = (\theta^{t-1}, \theta^t)$ のときの信念 $b_{\theta'}$ から，現在の状態における信念 b_{θ^t} を導ける．このとき，$v'(b_{\theta'}) = v(b_{\theta^t})$ が成立している．

3.3 プログラムインタフェース

このプログラムでは図 4 にあるように，割引因子（discount factor），成分ゲーム（stage game）の記述，$o(\boldsymbol{\omega} \mid \boldsymbol{a})$ で定義される観測構造（monitoring structure），すなわち，行動プロファイル $\boldsymbol{a}$ が与えられたときの私的シグナルプロファイル $\boldsymbol{\omega}$ を観測する確率の組合せ，そして 1 つの FSA を入力として用いる．以下に入力の例を紹介する．

```
# discount factor
discount: 0.9
# stage game (actions and payoff matrix)
actions: C D
PM:C:C: 1: 1
PM:D:C: 2:-1
PM:C:D:-1: 2
PM:D:D: 0: 0

# monitoring structure (observation and its
# probability)
observations: g b
O:g:g:C:C:0.97
O:b:g:C:C:0.01
O:g:b:C:C:0.01
O:b:b:C:C:0.01
...
# FSA description of Grim-trigger
states: R P
start: R
T:R:g:R
T:R:b:P
T:P:g:P
T:P:b:P
```

4. ノイズを含む観測付き繰返し囚人のジレンマ

本章ではまず，“ほぼ完全”（nearly-perfect）な観測構造を定義する．観測がほぼ完全であるとは，各プレイヤがつねに各期の相手の行動を十分に高い確率で完全に観測できる，すなわち，p が q と s よりも十分に大きく，$p = v$，$q = r = t = w$，$s = u = 1 - p - 2q$ となる場合とする．これは 2.3 節で述べたほぼ公的観測とは完全に異なることに注意されたい．ほぼ公的観測ではプレイヤがお互いに異なるシグナルを観測することはほとんどない．一方で，ほぼ完全観測では，プレイヤの行動によって，たとえば，(C,D) のとき，お互いに異なるシグナルを観測することが十分にありうる．また，ほぼ完全観測はきわめて自然な観測構造であるが，POMDP ソルバを使わなければ，十分な解析ができなかった．

以降ではとくに断らない限り，$x = 1$，$y = 1$，割引因子

$\delta = 0.9$ とする．$p + 2q + r = 1$ の制約の下，$p \in (1/2, 1)$ および $q \in (0, 1/4)$ 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．次に，我々のプログラムを用いて，GT，TFT，1-MP が SPFSE を構成するシグナルパラメータの範囲を明らかにする．

4.1 無限期罰則のトリガ戦略

本節では無限期罰則のトリガ戦略（GT）を表す FSA を検証する．2 人のプレイヤがともに GT に従って振る舞う場合，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

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

となり，これを解くことで，

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

を得る．

図 10 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 が大きいほど正確になる．つまり，相手が C（協力）/D（裏切り）を選ぶとき，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．たとえば，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，シグナルが正確で，その相関が強い領域で SPFSE を構成する．また，p が大きく q が小さくない場合，シグナルは正確だが，その相関が弱くなっている．ここでプレイヤ 1 が b を観測すると仮定すると，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分かる．

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．したがって，プレイヤ 1 は GT を無視して協力し続ける方がよい．一方で，p が比較的小さい場合，相手が b を観測する確率は大きくなる．したがって，プレイヤ 1 は裏切りから始める方がよい．GT の欠点は 1 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．たとえば，$p = 0.9$，$q = 0.01$，$\delta = 0.9$ 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 に従って行動を選ぶ場合はおよそ 5.31 にまで小さくなる．

4.2 しっぺ返しと 1-期相互処罰

GT よりも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 TFT をとる場合，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 5** にほぼ完全観測の下での TFT の積 FSA を示す．ここで，太線・細線・点線はそれぞれ p，q，s 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 s 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 5 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よい．したがって，ほぼ完全観測の下での TFT 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サブゲーム完全均衡を構成できない．そのうえ，TFT の組合せが実現する利得は非常に低くなる．これは図 5 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，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” [6] と知られている 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 6** 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 に再び遷移できる．プレイヤが RR 状態にいる（不変

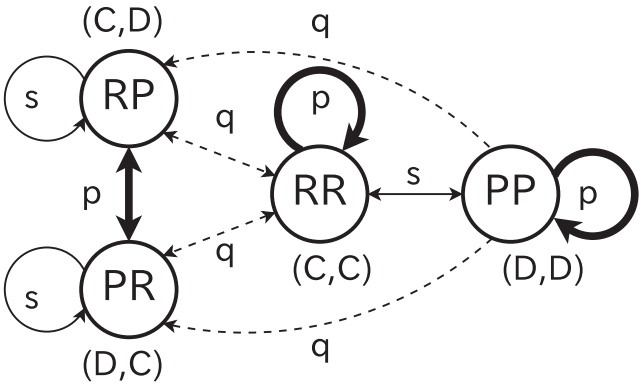

図 5 ほぼ完全観測下における TFT の積 FSA
Fig. 5 Joint FSA for TFT under nearly-perfect monitoring.

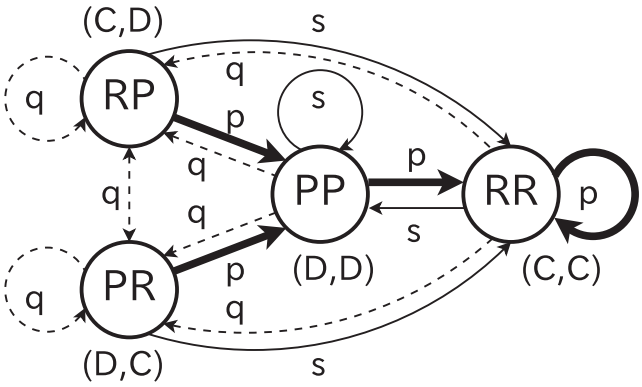

図 6 ほぼ完全観測下における 1-MP の積 FSA
Fig. 6 Joint FSA for 1-MP under nearly-perfect monitoring.

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の増加 x が次の期における利得の損失 1 と一致するため，将来の利得を割引く限り，1-MP は完全観測下でも SPFSE を構成できない[*2]．

5. k-期相互処罰

本章では，1-MP のアイデアを k-期相互処罰（k-MP）へと一般化する．この FSA では，プレイヤは最初に協力する．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 7 に 2-MP の，**図 8** に 3-MP の FSA を示す．2-MP に従って振る舞うプレイヤは状態 R にいるとき，相手が裏切る（シグナル b を受け取る）と，状態 P_1 に遷移する．ここから 2 回連続してシグナル b を受け取ると，状態 P_2 を経由して状態 R に戻る．3-MP の場合は 1 度 b を受け取ると，3 回連続して b を受け取ったとき，状態 P_2，P_3 を経由して状態 R に戻る．2-MP および 3-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，相手がつねに裏切る場合，2-MP は 3 回に 1 回は必ず協力し，3-MP は 4 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 9** は 2-MP の積 FSA を示している．簡単のため，最も大きい確率 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．また，FSA の初期状態，つまり最初にどの行動をとるかは結果に影響を与えやすい．しかし，本論文において k-MP が均衡を構成するか否かは，各プレイヤが最初にどの行動をとるかには依存しない．

[*2] $\frac{1+x}{1-\delta^2} < \frac{1}{1-\delta}$ となる場合に限り，1-MP は完全観測の下でサブゲーム完全均衡となる．

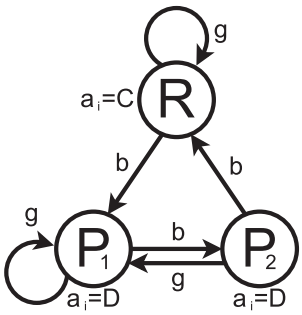

図 7 2-期相互処罰（2-MP）
Fig. 7 2-MP.

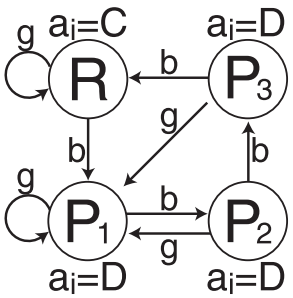

図 8 3-期相互処罰（3-MP）
Fig. 8 3-MP.

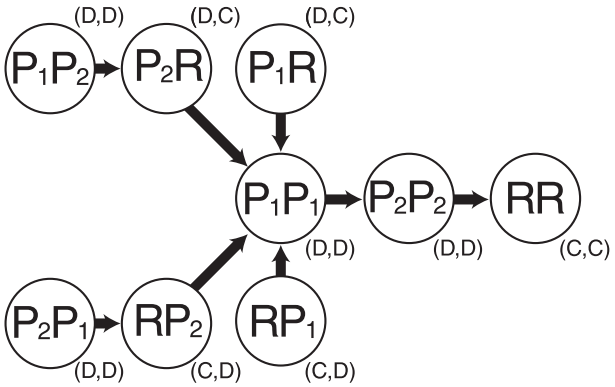

図 9 ほぼ完全観測下における 2-MP の積 FSA
Fig. 9 Joint FSA for 2-MP under nearly-perfect monitoring.

図 10 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 10 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．$q=0.04$ のとき，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 の方がシグナルの相関の強さの影響を受けやすいことである．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 10 には 3-MP

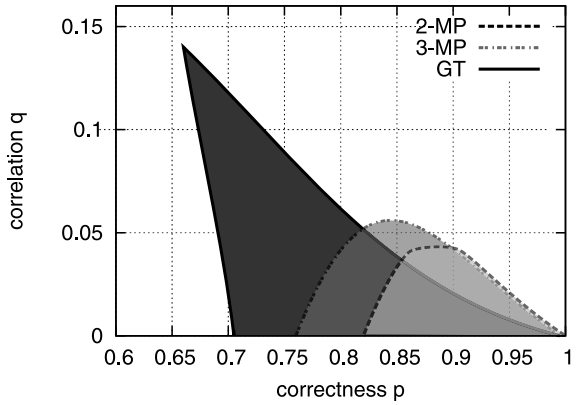

図 10 GT/2-MP/3-MP が SPFSE を構成するシグナルパラメータの領域（$p+2q \leq 1$）

Fig. 10 Range of signal parameters over which GT/2-MP/3-MP in an SPFSE. Note that feasible parameter space in $p+2q \leq 1$.

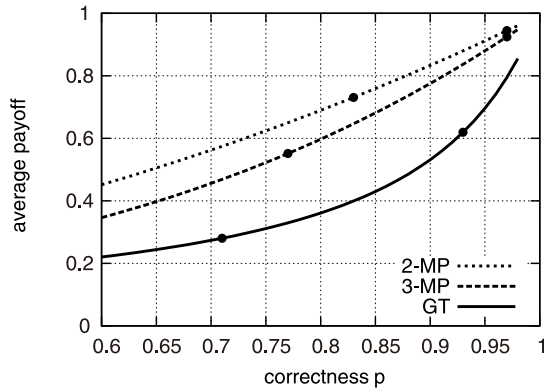

図 11 GT/2-MP/3-MP の期ごとの平均利得（$q = 0.01$）

Fig. 11 Average payoff per period of FSA ($q = 0.01$).

（図 8）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分かる．

図 11 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 軸は期ごとの平均利得を表す．また，平均利得が 1 になるのは相互協力がつね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 GT と 3-MP より高い平均利得を実現している．

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．

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 $|A|^{|\Theta|} \cdot |\Theta|^{|\Theta|\cdot|\Omega|} = 5{,}832$ 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 しかないことが分かった．

6. ランダムな私的シグナルを用いた拡張

本章では，これまで相手の行動に関するシグナルを 2 つだけ与えていたのに対し，ランダムな私的シグナルを 1 つ追加することを考える．この追加するシグナルは，(i) 利得に影響を与えない，(ii) プレイヤの行動についての情報を伝えない，(iii) 両方のプレイヤ間で強く相関している，といった性質を持つと仮定する．興味深いことに，このような相手の行動に関する情報を含まないシグナルを用いることでよりうまく協調的な振舞いを達成できる．具体的には，プレイヤは各成分ゲームの前に追加のシグナルを観測するというイベントが起こるかどうかによって，お互いの行動をよりうまく調整できる．ここで，両方のプレイヤが追加したシグナルを観測する確率を p'，どちらのプレイヤもこのシグナルを観測しない確率を s'，片方のプレイヤのみがこのシグナルを観測する確率を $(1-p'-s')/2$ と仮定する．p' は比較的小さく（ほとんど起こらないということはない），$(1-p'-s')/2$ は p' よりも非常に小さい，すなわち，1 人のプレイヤがシグナルを観測したとき，もう 1 人のプレイヤも同様に観測する確率が十分に高いと仮定する．

このとき，プレイヤは追加したシグナルをどのように利用する（もしくは無視する）だろうか？ ここで我々はパラメータ設定を GT が SPFSE を構成する範囲に仮定する．新しく定義したシグナルはプレイヤの効用/観測から完全に独立しているので，このシグナルを無視しても損をすることはない．したがって，従来の GT（このシグナルを無視する）は引き続き均衡を構成する．

これに対し，プレイヤ 2 は次の戦略に従うと仮定する：追加のシグナルを観測しない限りは GT を行うが，シグナルを観測したときは必ず状態 R へ遷移する．プレイヤ 2 がこの戦略を行うと仮定すると，プレイヤ 1 にとっては，プレイヤ 2 と同じ戦略をとることが最適反応となるだろう．これは，プレイヤ 1 がシグナルを観測すると十分高い確率でプレイヤ 2 もまたシグナルを観測し状態 R へ遷移するためである．GT が SPFSE を構成しているため，プレイヤ 2 の状態 R にいる確率が高い限りは，プレイヤ 1 にとっての最適反応は状態 R へ遷移することになる．したがって，この新しい戦略（GT-s と呼ぶ）は SPFSE を構成できる．さらに，これと同じ拡張を k-MP に適用した FSA を k-MP-s と呼ぶ．まとめると，ここで述べた追加シグナルは新しく繰返しゲームを再スタートさせるリセットボタンと見なすことができ，それによって罰をより軽減させることになる．

$p' = 0.88$，$s' = 0.1$，$(1-p'-s')/2 = 0.01$ としたときに，GT-s または 2-MP-s が SPFSE を構成するパラメータの範囲を確認する．**図 12** に GT/2-MP と GT-s/2-MP-s

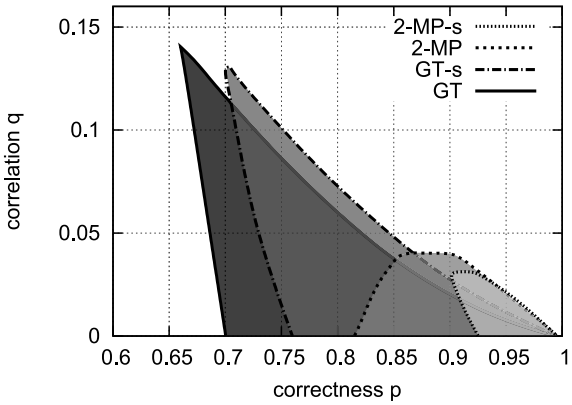

図 12 GT/2-MP と GT-s/2-MP-s が SPFSE を構成するシグナルパラメータの領域

Fig. 12 Range of signal parameters over which GT/2-MP and GT-s/2-MP-s are SPFSE.

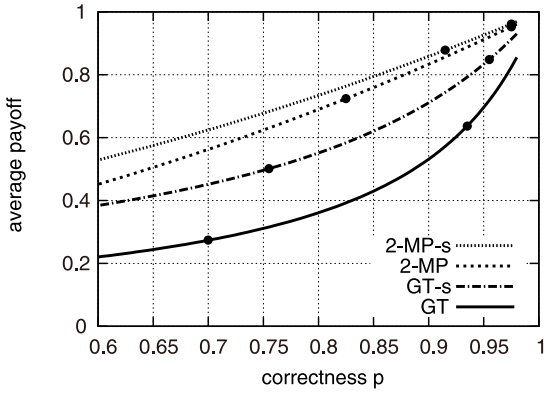

図 13 GT/2-MP と GT-s/2-MP-s の期ごとの平均利得 ($q = 0.01$)

Fig. 13 Average payoff per period of GT/2-MP and GT-s/2-MP-s ($q = 0.01$).

が SPFSE を構成するパラメータの範囲を示す．この図から，両方のプレイヤの正しいシグナルを観測する確率 p において，GT-s（2-MP-s）が SPFSE を構成する範囲は GT（2-MP）よりも小さいことが分かる．一方で GT のみに関していえば，片方のプレイヤが間違ったシグナルを観測する確率 q において，GT-s が SPFSE を構成する範囲は GT よりも大きい．**図 13** に毎期の平均利得を示す．この図からは，GT-s（2-MP-s）が SPFSE を構成する範囲は GT（2-MP）よりも小さいことが分かる．しかし，平均利得は追加のシグナルを導入する方が大きくなっている．

これと似たようなアイデアが文献 [2] に示されているが，ここでの追加シグナルは公的，つまりお互いのプレイヤがつねに同じシグナルを観測すると仮定されている．本章では，POMDP ソルバを用いることで，シグナルが私的，つまりお互いのプレイヤがつねに同じシグナルを観測するとは限らないケースを分析している．

7. 結論

本論文では不完全私的観測付き繰返しゲームにおける均衡を分析するプログラムを提案した．このようなゲームの分析は非常に困難な問題だと考えられていた．しかし，文献 [5] のアイデアをもとに，既存の POMDP ソルバを利用して，与えられた FSA が SPFSE を構成するかどうかを自動的に確認するプログラムを開発した．このプログラムを用いることで，ゲーム理論や人工知能/マルチエージェントの研究者を含む，POMDP の非専門家であってもソルバを用いて繰返しゲームの均衡を分析できるようになる．一般に POMDP は PSPACE 完全な問題であるため，提案プログラムで計算できるゲームの規模はそれほど大きくないが，私的観測付き繰返しゲームの均衡は有名な 2 人囚人のジレンマゲームでさえ，ほとんど知られていなかった．そこで，このプログラムの有用性を示すため，間違った観測を得る確率が比較的小さい場合における私的観測付き無限回繰返し囚人のジレンマの均衡を探索した．まず初めに，ノイズを含む観測が GT，TFT，1-MP（Pavlov）の振舞いに与える影響を吟味した．次に，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 GT と Pavlov 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．

そのうえで，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，GT 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 GT より高い利得を達成する FSA が 2-MP の他に存在しないことを確認した．

この結果は経済学分野で難しいとされてきた問題を，POMDP という情報処理の技術を用いて実際に解く方法を示すことで，POMDP の新しい適用領域を示せたと考えている．今後の課題として，プログラム全体の計算量の吟味や，3 人以上の非対称ゲームを計算するための高速なアルゴリズムの開発などがあげられる．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した現実に近い状況を，このプログラムを用いて分析していきたいと考えている．

参考文献

[1] Doshi, P. and Gmytrasiewicz, P.J.: On the Difficulty of Achieving Equilibrium in Interactive POMDPs, *Proc. 21st National Conference on Artificial Intelligence*, pp.1131–1136 (2006).
[2] Ellison, G.: Cooperation in the Prisoner's Dilemma with Anonymous Random Matching, *Review of Economic Studies*, Vol.61, No.3, pp.567–588 (1994).
[3] Hansen, E.A., Bernstein, D.S. and Zilberstein, S.: Dynamic Programming for Partially Observable Stochastic Games, *Proc. 19th National Conference on Artificial Intelligence*, pp.709–715 (2004).
[4] Kandori, M.: Game theory, *Repeated games*, pp.286–299, Palgrave macmillan (2010).
[5] Kandori, M. and Obara, I.: Towards a Belief-Based

Theory of Repeated Games with Private Monitoring: An Application of POMDP (2010). available from ⟨http://mkandori.web.fc2.com/papers/KObb10June4.pdf⟩.
[6] Kraines, D. and Kraines, V.: Pavlov and the prisoner's dilemma, *Theory and Decision*, Vol.26, pp.47–79 (1989).
[7] Mailath, G. and Samuelson, L.: *Repeated Games and Reputation*, Oxford University Press (2006).
[8] 松島 斉：ゲーム理論の新展開, 第4章「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] Nowak, M.: *Evolutionary Dynamics: Exploring the Equations of Life*, Harvard University Press (2006).
[10] 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).
[11] 岡田 章：ゲーム理論（新版）, 有斐閣 (2011).
[12] Russell, S. and Norvig, P.: *Artificial Intelligence: A Modern Approach*, *3rd Edition*, Prentice Hall (2009).
[13] 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).
[14] Tennenholtz, M. and Zohar, A.: Learning equilibria in repeated congestion games, *Proc. 8th International Conference on Autonomous Agents and Multiagent Systems*, pp.233–240 (2009).
[15] Wang, W., Chatterjee, M. and Kwiat, K.: Cooperation in Ad Hoc Networks with Noisy Channels, *Proc. 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, pp.1–9 (2009).
[16] 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).

ジョ ヨンジュン

2011年3月九州大学工学部電気情報工学科卒業．現在，九州大学大学院システム情報科学府修士課程在籍中．分散制約充足/最適化問題，ゲーム理論，意思決定論に関する研究に興味を持つ．

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．同年より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年より九州大学大学院システム情報科学研究院助教．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ノイズ付き繰返しゲーム等に興味を持つ．2012年情報処理学会論文賞，2011年FIT船井ベストペーパー賞．IEEE，人工知能学会各会員．

神取 道宏

1982年東京大学経済学部卒業．1989年スタンフォード大学経済学部Ph.D.，ペンシルバニア大学経済学部助教授，プリンストン大学経済学部助教授を経て，1999年より東京大学経済学研究科教授．繰返しゲーム，進化ゲーム理論の研究に従事するほか，制度設計や実験経済学にも興味を持つ．Ph.D.（経済学）．

小原 一郎

2001年ペンシルバニア大学にて経済学Ph.D.取得．2001年よりUCLAで助教授．2007年ミネソタ大学準教授を経て，2008年よりUCLA準教授．ゲーム理論，とくに繰返しゲームとメカニズム・デザインが専門．現在UCLA Center for Engineering Economics, Learning, and NetworksのCo-Directorを務める．

横尾 真 （フェロー）

1984年東京大学工学部電子工学科卒業．1986年同大学院修士課程修了．同年NTTに入社．1990～1991年ミシガン大学客員研究員．2004年より九州大学大学院システム情報科学研究院教授．マルチエージェントシステム，制約充足問題に関する研究に従事．エージェントの合意形成メカニズム，制約充足/分散制約充足等に興味を持つ．博士（工学）．1992年，2002年人工知能学会論文賞，1995年情報処理学会坂井記念特別賞，1999年，2005年人工知能学会全国大会優秀論文賞，2004年Association for Computing Machinery (ACM) Special Interest Group on Artificial Intelligence (SIGART) Autonomous Agent Research Award，2005年ソフトウェア科学会論文賞，2006年学士院学術奨励賞，2010年人工知能学会業績賞，International Foundation for Autonomous Agents and Multiagent Systems influential paper award受賞．人工知能学会，日本ソフトウェア科学会，電子情報通信学会，AAAI各会員．2011年AAAIフェロー．