4H1-3

# 学習あり繰り返し囚人のジレンマにおける協調行動の発生

Emergence of Cooperation in the Iterated Prisoner's Dilemma between Reinforcement Learners

鳥居 拓馬*1　日高 昇平*1　真隅 暁*1
Takuma Torii　Shohei Hidaka　Akira Masumi

*1北陸先端科学技術大学院大学 知識科学研究科
School of Knowledge Science, Japan Advanced Institute of Science and Technology

The iterated prisoner's dilemma (IPD) has been studied as a minimal model of cooperation. One of the key questions is the impact of learning—adaptively choosing actions based on past outcomes—on the dynamics of the game. Past studies have considered two distinct types of learning, the belief learning based on other opponents' behaviors and the reinforcement learning based on the rewards to its own actions. It has been known that the belief learners often converge into mutual defection. The potential impacts of the reinforcement learning, however, have not been fully understood yet. This study analyzed the reinforcement learners in the IPD as a finite Markov process, and showed its convergence into mutual cooperation when learning is sufficiently weighted. Our analysis revealed that this mutual cooperation emerged in a similar manner as in the IPD of the tit-for-tat strategies, which are known as the best heuristics. We discuss the effects of learning on mutual cooperation.

## 1. はじめに

社会における利害対立や協力関係を，合理的な推論をおこなう複数のプレイヤ間のゲームとして数理的に扱う枠組みとしてゲーム理論がある [Neumann & Morgenstern 44]．その代表例である囚人のジレンマは，複数のプレイヤのもっとも望ましい行動が協調行動であるにもかかわらず，裏切り行動が選ばれてしまう現象を捉えたゲームである．こうした基礎的なジレンマ構造は，社会性動物の利他的行動の発生，国際政治など，多様な文脈で見られ，囚人のジレンマにおける協調行動の発生に関する数理メカニズムの研究がなされてきた．本研究では，適応的に行動を選択するふたりのプレイヤ間の繰り返し囚人のジレンマ（Iterated Prisoner's Dilemma: IPD）をとりあげ，学習と協調の関係を調べる．

古典的な枠組みでは，各プレイヤが相手の行動やゲームの報酬などすべての情報をえられると仮定した信念学習が研究されてきた．この仮定の下では，仮想プレイ [Brown 51] や相手の行為の模倣（Tit-for-Tat: TFT）[Axelrod 84] などに代表される，相手の行動履歴にもとづく戦略が研究されてきた．

一方，相手の行動履歴を参照しない学習の枠組みとして，報酬にもとづき自身の行動を重みづける強化学習があげられる [Roth 95]．強化学習 [Sutton & Barto 98] の枠組みでは，相手の行動に関する情報をえられると仮定されておらず，したがって各プレイヤは自身のとりうる行動とそれに対する報酬のみからゲームの構造を間接的に推論することを要求される（他方で [Sandholm 96] のように相手の情報を参照できると仮定した強化学習もありえる）．学習するプレイヤ間のゲームの研究は比較的歴史が浅く，ゲームのダイナミクスに及ぼす学習の影響は解明すべき点が多く残されている [Sato 02]．

本研究では，ふたりの強化学習プレイヤ間の繰り返し囚人のジレンマを分析し，ゲームのダイナミクスに及ぼす学習の影響を調べた．ゲームの状態遷移を有限マルコフ過程として近似的に記述し分析した．

連絡先: 鳥居 拓馬，石川県能美市旭台 1-1，tak.torii@jaist.ac.jp

## 2. 繰り返し囚人のジレンマ

各プレイヤはふたつの選択肢，協調（Cooperate: C）もしくは裏切り（Defect: D），のうちいずれかの行動を決定する．ある時点 $t$ のプレイヤ $i \in \{1,2\}$ の行動を

$$x_{t,i} = \begin{cases} 0 & \text{if Cooperate} \\ 1 & \text{if Defect} \end{cases}$$

と記す．ふたりのプレイヤの行動（行動組み）は

$$x_t = (x_{t,1},\ x_{t,2}) \in \{0,1\} \times \{0,1\}$$

のいずれかとなる．ある時点 $t$ のプレイヤ $i$ と $j$（$i \neq j$）の行動に対応した報酬の組みは次の表に記されている．

| | | Cooperate ($x_{t,j}=0$) | Defect ($x_{t,j}=1$) |
|---|---|---|---|
| Cooperate | ($x_{t,i}=0$) | $r$(CC) | $r$(CD) |
| Defect | ($x_{t,i}=1$) | $r$(DC) | $r$(DD) |

このゲームが繰り返し囚人のジレンマであるためには，

$$2\,r(\mathrm{CC}) > r(\mathrm{CD}) + r(\mathrm{DC}) \quad \text{and}$$
$$r(\mathrm{DC}) > r(\mathrm{CC}) > r(\mathrm{DD}) > r(\mathrm{CD})$$

を満たす必要がある [Nowak 06]．本研究では，$r(\mathrm{CC}) = 3$，$r(\mathrm{CD}) = 0$，$r(\mathrm{DC}) = 5$，$r(\mathrm{DD}) = 1$ に固定して分析した．

### 2.1 強化学習

ある時点 $t$ までに，ふたりのプレイヤの行動の系列（履歴）

$$X_t = (x_{t-1},\ x_{t-2}, \ldots,\ x_{t-K})$$

が観察されたとしよう．この行動組みの系列に一意対応する報酬組みの系列が存在する．強化学習プレイヤ $i$ は自身への報酬の系列のみから，時点 $t$ での行動 $x_{t,i}$ を行動選択肢 $x \in \{0,1\}$ 上の確率分布（softmax 関数を用いて）

$$P_i(x) = \frac{\exp[\beta\, R_i(x)]}{\sum_y \exp[\beta\, R_i(y)]}$$

に従って確率的に決定する．ここで，

$$R_i(x) = \sum_{k=1}^{K} \alpha^k\, \delta(x, x_{t-k,i})\, r_{t-k,i}$$

は $x$ についての割引された累積報酬である．$\alpha$ と $\beta$ は定数である．$\alpha$ は過去の報酬の割引率と解釈される．$\alpha$ が大きいほど割引しなくなるので“ 保持率 ”といえる．$\beta$ が大きいほど累積報酬のより高い行動をより選択しやすくなり，$\beta = 0$ だと行動の選択は累積報酬に拠らずランダムになる．$\delta(x, y)$ は行動を判別する関数であり，

$$\delta(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$$

と定義される．明らかに，ある行動組みの系列 $X_t$ に一意対応する次時点の行動組み上の確率分布が存在する．

## 3. 有限マルコフ過程

$K$ 次マルコフ過程として記述するとしよう．IPD においては，プレイヤの数 $N = 2$，行動選択肢の数 $M = 2$ である．ふたりの行動の組みは $M^N$ 通りあるから，状態空間のサイズは $L = M^{NK}$，状態遷移行列 $\mathbf{Q}$ のサイズは $L \times L$ となる．行動組み $x_k = (x_{k,1},\ x_{k,2})$ は $M = 2$ 進数とみなして $1, \ldots,\ M^N$ の整数と一対一対応する．そうすると，行動組みの系列 $X = (x_1, \ldots, x_K)$ は $M^N = 4$ 進数とみなして $1, \ldots,\ L$ の整数と一対一対応する．ある行動組みの系列（状態）$X_t$ が与えられると，そこからの遷移先状態は $M^N$ 通りあり，それぞれの遷移確率は一意に定まる．状態遷移行列 $\mathbf{Q}$ は各列に $M^N$ 個の $\mathbf{Q}_{ij} > 0$ となる要素をもち，その総和は $\sum_j \mathbf{Q}_{ij} = 1$ となる．

Perron–Frobenius 定理によると，このような確率状態遷移行列は唯一の定常分布をもち，定常分布の計算は固有値問題に帰着される．したがって，$K$ 次マルコフ過程として記述されたモデルの解は

$$v_{t+1} = \mathbf{Q}\, v_t$$

の極限 $t \to \infty$ として，ただひとつ与えられる．

ある行動組みの系列 $X$ への滞在確率を $v_\infty(X)$ と記すことにすると，最終的に，ある行動組み $x$ （e.g. 協調行動）となる確率は周辺確率を計算することで与えられる：

$$\rho(x) = \sum_X \sum_{k=1}^{K} \frac{v_\infty(X)}{K}\, \delta(x, x_k)$$

ここで，$X = (x_1, \ldots, x_K)$ は状態空間のひとつの状態を表す．読みやすさのために，定常分布の周辺確率をそれぞれ $\rho$(CC)，$\rho$(CD)，$\rho$(DC)，$\rho$(DD) と記す．

## 4. 結果

強化学習 IPD の真の定常分布は $K \to \infty$ において与えられる．図 1 は今回の数値実験のなかで最大の $K = 7$ における

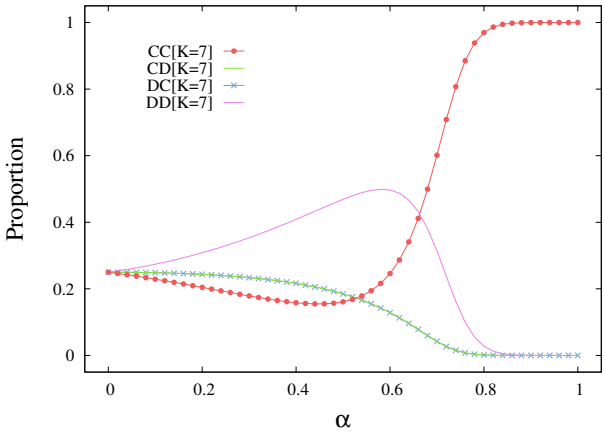


図 1: 保持率 $\alpha$ と周辺確率 $\rho$(CC)，$\rho$(CD)，$\rho$(DC)，$\rho$(DD) の関係．$\alpha$ が大きくなるほど CC 優位になる．$K = 7$，$\beta = 1.0$.

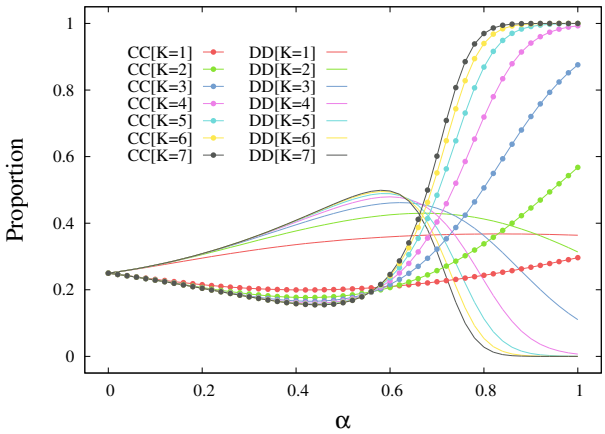


図 2: 保持率 $\alpha$ と周辺確率 $\rho$(CC)，$\rho$(DD) の関係．$K$ が大きくなるほど，より小さな $\alpha$ でも CC 優位になる．$K \in \{1, \ldots,\ 7\}$，$\beta = 1.0$．$\rho$(CD) と $\rho$(DC) は省略．

結果である．$\alpha = 0$ では理論どおり一様分布となる．$\alpha < 0.66$ くらいまでは DD 優位であるが，$\alpha > 0.66$ から転じて CC 優位となる．この結果は，学習に用いられる行動の履歴が相対的に長い場合に，両プレイヤの協調行動（CC）が高い頻度で現れることを示している．

図 2 は $K = 1 \to 7$ と変えたときの CC と DD のみの変化である．今回の利得行列では $K = 1$ のときのみ CC 優位とならない．図から，$K$ が大きくなるほど，より小さな $\alpha$ でも CC 優位になるとわかる．

次に，$K = 7$ に固定したうえで，異なる $\beta$ を比較する．図 3 から，$\beta$ が大きくなるほど，より小さな $\alpha$ でも CC 優位になるとわかる．図 1–3 は一貫して学習に用いる行動の履歴が長くなるほど協調行動の発生確率が高くなることを示している．

### 4.1 模倣的協調戦略の発現

パラメタの変化にともなう強化学習プレイヤの戦略の変化を評価するために，信念学習の代表的な戦略である Tit-for-Tat（TFT）戦略と Win-Stay, Lose-Shift（WSLS）戦略からの類似性を分析した．TFT プレイヤは過去 1 回の履歴をもとに相手の行動を模倣するという仕方で次の行動を決定するため，

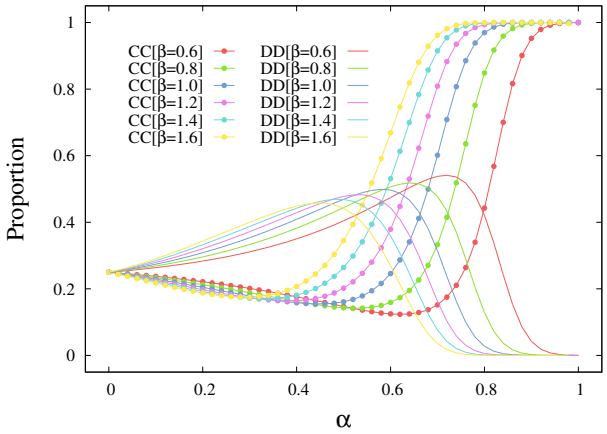


図 3: 保持率 $\alpha$ と周辺確率 $\rho$(CC), $\rho$(DD) の関係．$\beta$ が大きくなるほど，より小さな $\alpha$ でも CC 優位になる．$K = 7$, $\beta \in \{0.6, \ldots, 1.6\}$．$\rho$(CD) と $\rho$(DC) は省略．

TFT プレイヤ間のゲームは次のような遷移行列となる：

$$\begin{array}{c} \\ \text{CC} \\ \text{CD} \\ \text{DC} \\ \text{DD} \end{array} \begin{array}{c} \begin{array}{cccc} \text{CC} & \text{CD} & \text{DC} & \text{DD} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

WSLS は，一方が負けたとき（CD，DC）には負けたプレイヤが行動を変化させるので DD へ遷移し，両プレイヤが同じ行動を選択したとき（CC，DD）にはその状態（CC → CC, DD → DD）に遷移する．WSLS プレイヤ間のゲームは次のような遷移行列となる：

$$\begin{array}{c} \\ \text{CC} \\ \text{CD} \\ \text{DC} \\ \text{DD} \end{array} \begin{array}{c} \begin{array}{cccc} \text{CC} & \text{CD} & \text{DC} & \text{DD} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

TFT と WSLS それぞれの遷移行列の定常分布 $\rho$(CC), $\rho$(CD), $\rho$(DC), $\rho$(DD) と，各パラメタごとの強化学習プレイヤの定常分布の類似性を平均二乗誤差（MSE）で評価した．図 4 に保持率 $\alpha$ の関数として強化学習と 2 つの信念学習戦略（TFT と WSLS）との MSE を示す．この図から，$\alpha \approx 0.65$ を境に $\alpha$ が大きくなるにつれて，強化学習の状態遷移が TFT から WSLS へとますます類似してくることがわかる．また，定常分布から導出された強化学習の 2 次の遷移行列（図 5）はそれぞれ上記の（a）TFT 戦略，(b）TFT と WSLS 戦略の中間，(c）WSLS 戦略，と解釈できるパタンを示している．

TFT 戦略は，両者が TFT の場合は協調行動 CC を維持できるが，その場合にも一方が何らかのノイズにより誤った行動（D）をとると裏切り行為 DD に陥ることが指摘されている [Nowak 92]．これを踏まえ WSLS は TFT を改良し，確率的なノイズに対して頑強に協調関係を維持できる戦略として知られており，他の戦略よりも進化的に安定である条件が明らかになっている [Nowak 93]．図 4 の結果は，累積報酬の保持率が高くなるにつれ，強化学習が TFT からより確率的に頑強な WSLS に類似した行動選択をもつようになることを示している．この結果は，強化学習プレイヤが学習の結果として，

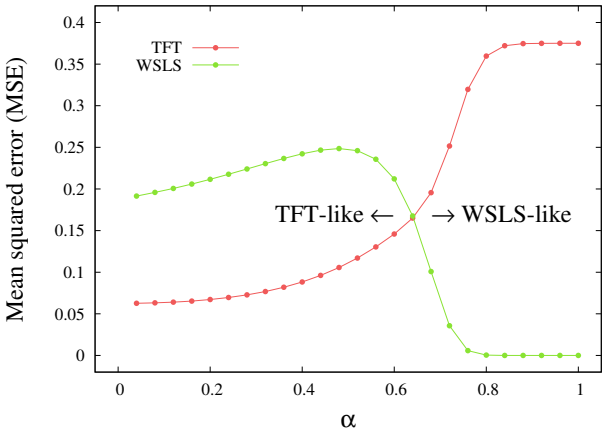


図 4: 強化学習モデルの周辺確率と TFT および WSLS モデルの周辺確率の誤差．$K = 7$, $\beta = 1.0$.

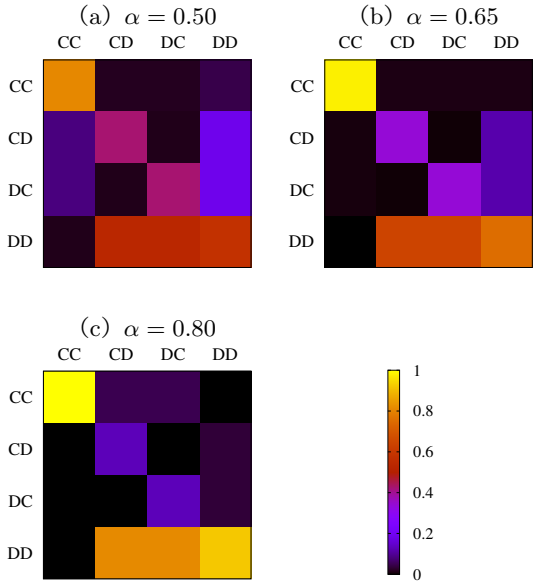


図 5: 強化学習モデルの 2 次の条件つき遷移行列．(a）TFT に近い場合（$\alpha = 0.50$）, (b）TFT と WSLS の中間的な場合（$\alpha = 0.65$）, (c）WSLS に近い場合（$\alpha = 0.80$）.

保持率に応じてより頑強な互恵的戦略を学習することを示唆している．

## 5. 議論

古典的囚人のジレンマおよびその流れの IPD 研究（e.g. 信念学習）によると，協調行動は稀れにしか安定して発生しない．さらに，信念学習ではより長い行動履歴を利用するほど（$\alpha \to 1.0$, $K \to 7$）裏切り行動が優位になると知られている [Axelrod 84]．これとは逆説的に，強化学習では $\alpha \to 1.0$, $K \to 7$ にともなって協調行動がますます優位になることが本研究から示された．この逆説的な協調行動の発生を説明するためにおこなった TFT および WSLS 戦略との比較から，強化学習は明示的な戦略が与えられていないにも関わらず，この 2 つの戦略を学習の帰結として形成している可能性が示された．

本研究で分析した強化学習では，相手の行動は直接的に参照できない．しかし，各プレイヤへの報酬は相手の行動の関数

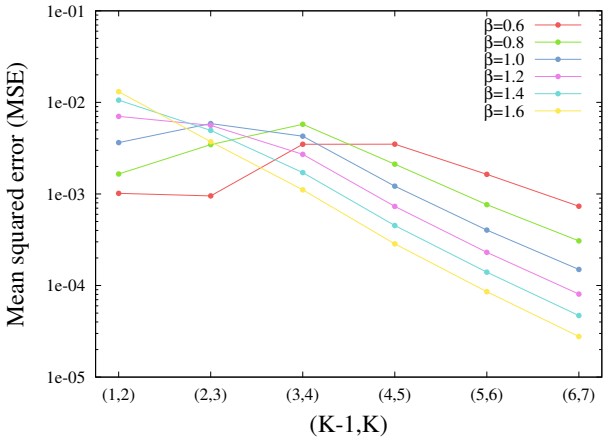


図 6: $(K-1, K)$ 対ごとの平均二乗誤差．$4 \leq K \leq 7$ においては指数的に減少している．

であるため，十分に長い過去の系列から相手の行動を間接的に学習することが可能である．より長い過去の履歴を参照した場合，自身の裏切り行為は，相手の裏切り行為へとつながり，長期的には自身の損失に繋がる．こうした推論は，報酬からの学習という形で強化学習により表現され，WSLS 戦略に近い行動戦略の頻度が高まると考えられる．

### 5.1 有限マルコフ過程による近似精度

本研究では有限マルコフ過程による分析を次数 $K \leq 7$ についておこなった．しかし，実際には $K \to \infty$ の状態遷移は計算不可能であるため，有限次数での打ち切りによる誤差が発生する．したがって，本研究で採用した有限マルコフ過程による近似誤差を評価するため，行動組み系列の長さ $K$ に応じて誤差がどのように変化するか調べた．図 6 は $K \leq 7$ の各次数の隣接対について，$\alpha \in \{0,\ 0.02, \ldots,\ 1\}$ について周辺確率 $\rho$(CC)，$\rho$(CD)，$\rho$(DC)，$\rho$(DD) の平均二乗誤差（MSE）を計算した結果である．IPD において，誤差は十分に大きな $K$ について，指数的に減少することがわかる．また，本研究では報告していないが，エージェント・シミュレーションと有限マルコフ過程による計算とは $K \geq 8, \alpha < 0.7$ でも標本誤差程度の誤差しかないことがわかっている．

特異的なパラメタ領域 $\alpha \approx 1.0$ では，エージェント・シミュレーション [Sandholm 96]，有限マルコフ過程による近似はともに精度が急激に落ちる．このようなケースにおいてどのような分析を行うべきかなど，いくつか技術的な課題が残されている．

## 6. 結論

適応的に行動を変化させる学習プレイヤ同士のゲームのひとつとして，強化学習プレイヤの繰り返し囚人のジレンマ（IPD）をとりあげ，有限マルコフ過程として分析した．固定戦略や信念学習プレイヤの IPD では協調行動が発生しにくいこと，行動履歴が長くなるほど裏切り行動が優位になることが知られているが，強化学習プレイヤの IPD では行動履歴が長くなるほど協調行動が優位となるとわかった．

## 参考文献

[Axelrod 84] Axelrod, R.: The Evolution of Cooperation, Basic Books, New York (1984).

[Brown 51] Brown, G.W.: Iterative solution of games by fictitious play. In: Koopmans, T.C. (ed) Activity Analysis of Production and Allocation, John Wiley & Sons, New York, 374–376 (1951).

[Neumann & Morgenstern 44] von Neumann, J. & Morgenstern, O.: Theory of Games and Economic Behavior, Princeton University Press (1944).

[Nowak 90] Nowak, M.A.: Stochastic strategies in the prisoner's dilemma. Theoretical Population Biology, 38, 93–112 (1990).

[Nowak 92] Nowak, M.A. & Sigmund, K.: Tit-for-tat in heterogeneous populations. Nature, 355, 250–253 (1992).

[Nowak 93] Nowak, M.A. & Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. Nature, 364, 56–58 (1993).

[Nowak 06] Nowak, M.A.: Evolutionary Dynamics. The Belknap Press of Harvard University Press (2006).

[Roth 95] Roth, A.E. & Erev, I.: Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. Games and Economic Behavior, 8, 164–212 (1995).

[Sandholm 96] Sandholm, T.W. & Crites, R.H.: Multiagent reinforcement learning in the iterated prisoner's dilemma. Biosystems, 37(1–2), 147–166 (1996).

[Sato 02] Sato, Y., Akiyama, E., & Farmer, J.D.: Chaos in learning a simple two-person game. Proceedings of the National Academy of Sciences, 99:7, 4748–4751 (2002).

[Sutton & Barto 98] Sutton, R.S. & Barto, A.G.: Reinforcement Learning: An Introduction, MIT Press (1998).