JAIST Repository
https://dspace.jaist.ac.jp/

| | |
|---|---|
| Title | 不確実な状況における利己的な学習主体の相互協調 |
| Author(s) | 鳥居，拓馬 |
| Citation | |
| Issue Date | 2016-03 |
| Type | Thesis or Dissertation |
| Text version | ETD |
| URL | http://hdl.handle.net/10119/13511 |
| Rights | |
| Description | Supervisor:橋本　敬，知識科学研究科，博士 |

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

博　士　論　文

# 不確実な状況における利己的な学習主体の相互協調

鳥居　拓馬

主指導教官　　橋本　敬

北陸先端科学技術大学院大学

知識科学研究科

平成 28 年 3 月

# Mutual cooperation between greedy learners under uncertainty

Student No: 1060005
Name: Takuma Torii

Cooperation is a common form of social interaction. The problem of cooperation is a conflict between individual rationality and collective rationality: selfish players maximizing their own profit obtain a smallest profit as a collective. Since in many situations in nature or society behaving selfishly is apparently a profitable option, how organisms solve the problem of cooperation is a long-lasting question. The problem was formulated as Prisoner's Dilemma in game theory, and various theoretical and empirical studies have shown possibilities of cooperation.

In game theory, a rational player takes a payoff-maximizing strategy (or action). A payoff to each player is a function of strategies of all players. A profile of strategies is called a Nash equilibrium if no one can improve his payoff by unilaterally changing his strategy. Prisoner's Dilemma is a game, in which each of two players has two options: Cooperate or Defect. It has been proved that the only Nash equilibrium in one-shot Prisoner's Dilemma is mutual defection. Iterated Prisoner's Dilemma (IPD) is its extensive-form variant. It has been proved that mutual cooperation can be a Nash equilibrium as well as mutual defection if both players consider sufficiently long-term future. One of the prominent findings in this area is the so-called tit-for-tat (TFT) or reciprocal strategy, whose behavioral rule is: if you cooperate (defect) I will cooperate (defect). Many theoretical studies has provided evidence supporting TFT for cooperation in some context, however, for what objective one acquires TFT-like behaviors is an unanswered question.

A key idea in recent game theory is bounded rationality: a decision maker has to choose an action based on limited information and restricted cognitive resources. Learning is a means to overcome uncertainty arising from incompleteness of information. Some psychological studies have shown that human cooperation is observed more frequency under uncertainty. Chapter 1 and 2 contains the background and reviews regarding the problem of cooperation in psychology and game theory.

This thesis aims at proposing a theoretical description for the problem of cooperation between two learning players under uncertainty. Concretely, the thesis aims at showing some conditions for mutual cooperation and what the learning players can maximize to establish mutual cooperation. Led by findings from psychology and game theory, in this thesis, the problem is formulated as Iterated Prisoner's Dilemma under uncertainty, where one can only get feedbacks to him in response to his actions. That is, no one can get information about the payoff matrix and related to the opponent. Reinforcement learning is a mechanism that can maximize its total profit only based on feedbacks (payoffs) in response to its actions. There are many evidence that reinforcement learning can solve various real world problems in uncertain environments.

In this thesis, we showed some conditions for mutual cooperation that can be established between selfish, reinforcement learners who attempt to maximize their own profit under uncertainty. Mutual cooperation is observed almost surely if both players make decisions based on sufficiently long-term experience. From a detailed analysis, TFT-like behaviors are observed during mutual cooperation between reinforcement learners. This finding gives TFT a position as a by-product of selfish/greedy learning. Chapter 3 includes the findings above.

Further conditions are investigated by approximating IPD of reinforcement learners. Using this approximated model, we formally studied several properties of the game, including payoff-related conditions for mutual cooperation. Based on the findings from the approximated model, we derived a payoff matrix that dramatically improve mutual cooperation between reinforcement learners. This can be interpreted as a practical mechanism for cooperation. The approximated model was studied in Chapter 4.

Chapter 5 contains a framework for future studies, which allows us to study reinforcement learning strategies within a more generalized class, although there are technical issues. In this thesis, we studied 1st-order strategy class, and discussed future directions.

Combining all the findings in this thesis and previous studies, a theoretical description for the problem of cooperation between learning players is argued in Chapter 6. It states that mutual cooperation can be estab-

lished under uncertainty if both selfish players learn to maximize their own profit from their long-term trial-and-error experience.

The problem of cooperation is a special case of the free-rider problem and/or shared resource problem, which are more common in real social situations. For example, one reported that leaving the problem of cooperation unsolved declines the performance of team members. The problem of cooperation can be one of the common barriers that impede performance of organizational activity, such as organizational knowledge creation. The findings in this thesis will provide a clue to resolve conflict situations in real social interactions. Contrary to a common belief about uncertainty, our finding suggests the possibility that uncertainty regarding information, especially conflict relationship, might improve mutual cooperation, if participants can learn from their action-feedback experience.

**Keywords:** Cooperation, Reinforcement learning, Game theory, Uncertainty, Prisoner's dilemma

# 目 次


**1 序論：協調問題 8**

1.1 協調問題への取り組み . . . . . 8
1.2 協調行動の説明とは . . . . . 11
1.3 本論文の主題 . . . . . 12
1.4 本論文の構成 . . . . . 14

**2 先行研究：相互協調の可能性 15**

2.1 本章の目的 . . . . . 15
2.2 囚人のジレンマ . . . . . 15
2.3 繰り返し囚人のジレンマ . . . . . 17
2.3.1 有限回繰り返し囚人のジレンマ（後方帰納） . . . . 19
2.3.2 無限回繰り返し囚人のジレンマ（素朴定理） . . . . 19
2.4 進化ゲーム理論 . . . . . 20
2.5 メタゲーム理論 . . . . . 22
2.6 学習主体のゲーム . . . . . 23
2.7 ゲーム構造の改変 . . . . . 25
2.8 ヒトの限定合理的な意思決定 . . . . . 26
2.9 まとめ . . . . . 27

**3 強化学習戦略による協調問題の解決 29**

3.1 本章の目的 . . . . . 29
3.2 定式化 . . . . . 29
3.2.1 １次戦略 . . . . . 30
3.2.2 強化学習戦略 . . . . . 30
3.3 分析方法 . . . . . 33
3.3.1 有限状態マルコフ過程 . . . . . 33
3.3.2 周辺確率 . . . . . 34
3.4 強化学習戦略による相互協調 . . . . . 34
3.4.1 記憶保持率の分析 . . . . . 35

3.5 1次戦略との比較分析 . . . . . 36
3.5.1 定常周辺分布の分析 . . . . . 37
3.5.2 強化学習戦略の1次戦略による近似 . . . . . 38
3.6 異なる利得行列を用いた分析 . . . . . 43
3.7 強化学習戦略の1次戦略に対する振る舞い . . . . . 45
3.7.1 まとめ . . . . . 46
3.8 強化学習戦略の“進化” . . . . . 47
3.9 議論 . . . . . 50
3.9.1 まとめ . . . . . 50
3.9.2 他のアプローチとの関係 . . . . . 50
3.9.3 不確実性のある意思決定 . . . . . 51
3.9.4 強化学習を用いたIPD研究との関連性 . . . . . 52
3.9.5 学習の結果としてのしっぺ返し行動の発現 . . . . . 52
3.9.6 実験的な知見との関連性 . . . . . 53
3.9.7 今後の課題 . . . . . 53

**4 強化学習戦略を扱うゲームのベクトル場近似 55**
4.1 本章の目的 . . . . . 55
4.2 学習主体のゲームの近似 . . . . . 55
4.3 定式化 . . . . . 56
4.3.1 強化学習のベクトル場近似 . . . . . 56
4.3.2 固定点，ヌルクライン，安定性 . . . . . 58
4.3.3 $\Gamma_i$ の性質と略記 . . . . . 59
4.4 ベクトル場の可視化 . . . . . 59
4.5 特殊ケース . . . . . 64
4.5.1 特殊ケースの可視化 . . . . . 65
4.5.2 特殊ケースのDD優位解 . . . . . 67
4.5.3 特殊ケースのCC優位解 . . . . . 67
4.5.4 特殊ケースの解の個数と安定性 . . . . . 67
4.6 一般ケース . . . . . 71
4.6.1 一般ケースのDD優位解 . . . . . 71
4.6.2 一般ケースのCC優位解 . . . . . 71
4.6.3 一般ケースの解の個数と安定性 . . . . . 72
4.7 CC優位解へ到達しやすい利得行列 . . . . . 76
4.8 無条件報酬の影響 . . . . . 78
4.8.1 CC優位解の存在条件（無条件報酬あり） . . . . . 79

4.9 議論 . . . . . 80
4.9.1 均衡解の存在条件とその個数 . . . . . 80
4.9.2 協調行動の説明との関係 . . . . . 80
4.9.3 有限マルコフ過程との相補性 . . . . . 81
4.9.4 先行研究の近似モデルとの関係 . . . . . 82

**5 高次マルコフ戦略と今後の課題 83**
5.1 本章の目的 . . . . . 83
5.2 マルコフ戦略の無限回ゲーム . . . . . 83
5.2.1 有限マルコフ過程と定常分布 . . . . . 85
5.3 定式化 . . . . . 85
5.3.1 1次戦略 . . . . . 85
5.3.2 $K$ 次戦略 . . . . . 87
5.3.3 局所 Nash 均衡 . . . . . 87
5.4 1次戦略の局所 Nash 均衡 . . . . . 88
5.4.1 局所 Nash 均衡の数理解析 . . . . . 88
5.4.2 局所 Nash 均衡の数値検証 . . . . . 93
5.5 議論 . . . . . 95
5.5.1 1次戦略とその背景にある理論の関係 . . . . . 95
5.5.2 高次戦略を理解するための1次戦略 . . . . . 95

**6 総合考察 97**
6.1 本論文の主題とモデル . . . . . 97
6.2 協調問題の解決可能性 . . . . . 98
6.2.1 情報の次元と時間の次元 . . . . . 98
6.2.2 副産物としての返報性 . . . . . 100
6.2.3 協調促進メカニズム . . . . . 101
6.2.4 協調の計算論に向けて . . . . . 101
6.3 実社会への応用可能性 . . . . . 103

**7 結論 105**
7.1 まとめ . . . . . 105
7.2 不確実な状況における学習主体の相互協調 . . . . . 106
7.3 今後の展望 . . . . . 107

**A 囚人のジレンマの一般型 109**
A.1 共有地の悲劇 . . . . . 109
A.1.1 強化学習戦略と共有地の悲劇 . . . . . 110
A.2 公共財ゲーム . . . . . 112
A.2.1 強化学習戦略と公共財ゲーム . . . . . 113

# 1　序論：協調問題

複数主体の利害が対立する社会的問題状況では，各個人の意思決定がその寄せ集めとして集団的に望ましい帰結を導くかどうかが主要な問いのひとつである．ゲーム理論は社会的問題状況における利己的な主体の意思決定を分析する枠組みであり，個人的に望ましい帰結と集団的に望ましい帰結の対立は「協調問題」あるいは「囚人のジレンマ」として定式化され，分析されてきた．

## 1.1　協調問題への取り組み

協調は自然や社会のうちに数多く見られるが，いかにして協調問題を解決しているのだろうか．協調問題とは「各主体が個人として最も望ましい状態を追求すると，個人の行為の寄せ集めとして，集団として最も望ましい状態が実現されない」という個人最適と集団最適の対立をいう．協調問題にともなう利害対立の状況では，自己の利益を増やすことは他者の利益を減らすことに繋がり，双方の利益を同時に最大化することはできない．また，双方が自己利益のみを追求する行動をとる場合，互いに利益をえられない（損失を受ける）など集団的に望ましくない結果となる．こうした状況では，双方が互いに妥協した行動をとることで集団的に望ましい解決を達成しうる．しかし，こちらの妥協を逆手にとって相手に利益を搾取される可能性が残る．したがって，相手の行動に関して確証をもてない場合，自己利益を追求することが個人として最適な行動といえる．双方がこのように意思決定するならば，やはり集団としては望ましくない帰結がもたらされてしまう．

この予想に反して，現実の自然や社会では，各個体は互いに助け合うよう行動しているという証拠が提案されており，多くの研究者は自然や社会では持続的な協調が実現されていると考えている（レビュー論文として [1, 18, 44, 62, 64]）．また，自然では一見すると他者の利益となる利他的な行動が観察されている．もしそうならば，なぜ／何のために個

人は損失を負ってまで他者の利益となる行動をとるのだろうか．この予想と観察とのズレは数多くの協調問題に関する研究を動機づけてきた．

協調問題は，ゲーム理論において囚人のジレンマとして定式化され，複数分野の経験的知見とともに，協調行動の理論の発展を牽引してきた．ゲーム理論では，自己利益を最大化するよう意思決定を行う合理的プレイヤを想定する [63]．合理的プレイヤは自己利益を最大化する戦略（選択肢）を選び，全プレイヤの戦略の関数として各プレイヤへの利得が定まる．どのプレイヤも独立に戦略を変えても利得を高められないとき，その戦略の組みを Nash 均衡（あるいは単に均衡）という．古典的なゲーム理論は，完全情報の下で合理的選択を行う完全合理的プレイヤを想定し，囚人のジレンマなどのゲームを調べてきた．囚人のジレンマでは，各プレイヤは「協調」か「裏切」かのどちらかを選択する．集団最適は相互協調であるが，個人最適は他個体の行動に依らず裏切であるため，最も望ましくない相互裏切に陥ってしまう．相互裏切は 2 人の完全合理的プレイヤを想定したときの唯一の均衡であり，相互協調はありえないことが理論的に示されている．それ以来，どのような設定ならば，相互協調が実現されるかが研究されてきた[1]．繰り返し囚人のジレンマは「何度も繰り返される囚人のジレンマをどうプレイするか」を戦略とするゲームであり，完全合理的プレイヤの間では，有限回繰り返しの場合は相互裏切しかないが，無限回繰り返しの場合は相互協調と相互裏切の双方が可能であると理論的に示されている [23]．

進化生物学では繰り返し囚人のジレンマを題材とした進化ゲーム理論を用いて協調問題が研究されている．自然選択は広義にはその環境において個体の生存率・生殖率を最大化するような個体の行動を形成する．すなわち，自然選択は利己的な個体を好み，合理的な利己的行動を個体にもたらすと考えられている．進化ゲーム理論では，生物の集団を戦略の集団と捉え，その頻度分布が自然選択によりどう変化するか，自然選択の結果として集団を支配する合理的戦略はどのようなものかを問題とする．協調問題の困難は，協調する相手に対して裏切で応えることで，相手からより高い利益を搾取できる点にある．こうした搾取（タダ乗り）の誘因がある状況では，協調は維持されず崩壊し，裏切が容易に蔓延して

[1]何をもって相互協調が実現できたと定義するかはモデルの設定により異なる．強い定義は「相互協調が Nash 均衡となる」というものだろう．弱い定義は「相互協調の状態が繰り返される持続期間」や「相互協調の状態が生じる確率」と関連づけてなされる．本学位論文では「相互協調の状態が生じる確率」により定義し，相互協調の状態が十分に高い確率で生じるときを相互協調の実現と呼ぶ．

しまう．したがって，協調が持続的に実現されている場合には，何らかのメカニズムによって，こうした誘因が抑制される必要がある．自然選択の適用範囲に関するメカニズム（血縁選択，集団選択，空間選択）のほか，個体の行動原理に関するメカニズム（直接返報性，間接返報性）[2]が提唱されている [43]．返報性とは「君が協調（裏切）するなら僕も協調（裏切）する」という行動原理をいい，返報性に基づく行動戦略は「しっぺ返し戦略」と呼ばれている．しっぺ返し戦略（Tit-for-Tat）は，協調しながらも侵略や搾取に対抗できる，協調と裏切を兼ね備えた行動戦略 [3] であり，現実の協調行動に関して示唆を与えている．

現実の生物の行動は，進化のほかに，学習によっても組織化されうる．限定された情報のもとで合理的選択を行う限定合理的プレイヤへの関心からも，学習は調べられている．限定合理性とは，狭義には Simon [54] による最大化に代わる概念としての満足化を意味するが，広義には主体の認知資源や情報処理の限界のために局所最大化しかできないことを意味する．学習は，広義の意味での限界のなかで局所最大化を実現する手続きと位置づけできる．学習を扱った理論研究では，利得行列や相手の行動など，意思決定に利用できる情報の違いおよび学習機構の違いによって相互協調を実現できる場合があることが示されている（詳細は 2.6 節）．協調問題を扱った心理学的な研究から，ヒトの意思決定は必ずしもゲーム理論の想定する完全合理性をみたさないことが指摘されている．心理学の研究から，ヒト被験者の場合では，完全合理性を想定した理論よりも頻繁に協調が確認されている．また，繰り返し回数や利得に関する情報の不確実性が被験者の学習速度やゲームの結果に影響を与えることが示されている（詳細は 2.8 節）．これらの知見を踏まえると，完全合理的な意思決定により生じる個人と集団の間のジレンマが，ゲームに関する部分的な情報とそれを補う学習・推論に基づく限定合理的な意思決定では生じない可能性がある．

これまで見てきたように，協調問題に関する研究は「生物は協調している」という観察と「協調は実現しえない」という予想の間の矛盾に動機づけられている．協調問題が囚人のジレンマとして定式化されて以来，完全合理性，繰り返し回数，自然選択，限定合理性など，自然や社会から着想をえて，協調の実現可能性（ジレンマの解消可能性）について研究されてきた．これらの研究は，囚人のジレンマにおける相互協調の条

[2]ある利害をもたらす行為を他者から受けたとき，同じ行為を直接の利害関係にある当事者に対して行うとき，直接返報性という．他方，直接の利害関係にない第三者に対して行うとき，間接返報性という．ボランティアや八つ当たりは間接返報的といえる．

件を通して，現実の協調行動を説明しようとする．進化ゲーム理論では，進化シミュレーションにおいて集団内でより高い期待利得をえる戦略を探求してきたが，ある戦略が相互協調を実現できたとしてもその理由は「その集団において高い適応度をもたらすから」となる [40]．こうした進化論的目的による行動の説明はひとつの方法だが，他に，主体を情報処理システムとみなし，その計算目標すなわち計算論的目的による行動の説明がありうる．

## 1.2 協調行動の説明とは

認知科学の黎明期において，Marr [35] は認知主体の行動に関する説明を３つに区別している．認知科学とはある認知過程を「計算」として理解するパラダイムであり，とりわけシステム内部での情報表現と情報処理を重視する．この分類のなかで，Marr は，計算論のレベル，表現とアルゴリズムのレベル，そしてハードウェアのレベルを論じている．計算論のレベルは計算の目的，なぜ／何のためにその行動は生みだされるかを問う．換言すれば，目的関数，制約条件，ロジックなどが計算論のレベルに含まれる．表現とアルゴリズムのレベルは計算論の実現方法を問い，計算の目的やロジックを実現しうる情報処理（アルゴリズム）や情報表現を問う．最後に，ハードウェアのレベルは表現とアルゴリズムの物理的な実現方法を問う．通常，ある計算論は複数のアルゴリズムで実現でき，またあるアルゴリズムは複数のハードウェアで実現できる．

各レベルは認知過程の理解に関して相補的な関係にある．例えば，ある計算論のレベルでの仮説の正しさは何らかのアルゴリズムによる実装を通して検証できる．また逆に，ある状態を実現できるアルゴリズムを先に見つけ，計算論のレベルに対する仮説を立てることもできる．

例として，整数列の並び替え（ソーティング）問題を考える．この問題を扱う情報処理システムの目的は「整数列を昇順（降順）に並べる」である．実際，この計算目的は複数のアルゴリズムで解けることが知られている（クイックソートなど）．また，個々のアルゴリズムは通常デジタルコンピュータで実現されるが，ハードウェアとしての人間がアルゴリズム通りに手作業で並び替えても実現できる．

von Neumann and Morgenstern [63] によれば，社会的意思決定（あるいはゲーム的問題状況）の特徴は各個人の利益がその個人には（完全には）制御できない他者の行動に相互依存している，という点にある．[63]

はこれを「各個人の目的関数が自己の変数に加えて他者の変数を含む」と表現し，他者の変数を含まない最大化問題と対比させている．社会的意思決定は複数主体の認知過程すなわち「計算」が相互作用する状況と捉えられる．したがって，ゲーム的問題状況における計算論のレベルの説明はこうした「計算」の相互作用を記述し，個別主体の観点から自己と他者の変数がどのように変化し達成すべき集団的帰結を導くかを特徴づけるものとなるだろう．このような抽象的で複雑な問題状況を記述するためには，まずは，どのような最小要件のもとでその問題が解決可能かを表現とアルゴリズムのレベルで示すことが重要だと考えられる．そこでえられた知見は計算論のレベルの記述を定式化する手がかりとなるだろう．

## 1.3 本論文の主題

本学位論文では，近年の心理学的な知見をもとに，情報の不確実性を補う学習による問題解決を支持する立場から，協調問題を「利害関係や他者の情報に関する不確実性の高い問題状況」と捉え，学習主体による不確実な協調問題の解決を扱う．

経済学においては，不確実性と類似した用語にリスクがある．この区別を提唱した Frank Knight の思想解説 [67] によれば，リスクは事象の生起を統計的に予測可能な状態を，他方，不確実性とは何らかの理由で統計的に予測困難な状態を意味する．その理由の１つに，希な現象のため，統計的に十分なデータの収集困難がある．たとえば，金融経済学ではリターンの分布として正規分布を仮定し，分散をリスクの指標とする考えがあると同時に，金融ショックなどは正規分布を逸脱した挙動を示し，そのため不確実性をもつとされる．

本学位論文では，個別主体は問題状況に関して限定的な情報しかもたず，さらに複数の主体が相互依存しながら同時に学習する状況を扱う．こうした状況では，学習主体の直面する問題・環境（他者や利得行列を含む）は自分と同じ速さで動的に変化するため，統計的な扱いが困難であると考えられる．本学位論文では，学習主体の観点から，問題・環境のもつ動的な性質のために，ある事象の生起確率を一般的には経験的に推定困難な状態を指す意味で「不確実性」という用語を用いる．ここで，情報の不確実な問題状況とはより具体的には以下のような状況をいう．

複数の主体が関与する社会的問題状況のなかには，その社会的問題の

全体像（関係者や利害関係など）を個別主体が正確に把握できないものがある[3]．この場合，個別主体は，直面する「部分的な問題」（実際にはある社会的問題の一部分）を問題の全体像だと思い込み[4]，部分的な問題に関する自己利益の最大化を試みることになる．このような部分的な問題状況では，集団全体としては１つの問題に直面しているにもかかわらず，個別主体の観察可能な情報のみでは全体像を同定できないため，個別主体の目線からみればアクションに応じて返されるフィードバックが一意に定まらない．このように部分観察的な情報しか利用できないために不確実性をもったフィードバックしかえられない問題状況を，本学位論文では，不確実な問題状況という．個別主体はフィードバックの不確実性に対処する必要があるが，不確実な問題状況に繰り返し直面する場合，アクション・フィードバックの試行錯誤的な経験から学習することが考えられる．

本学位論文では，このような不確実な協調問題を扱い，「計算」による行動の理解を目的とする認知科学の立場から，学習主体による相互協調の計算論的理解を目指す．ここで，計算論的理解とは，Marr の分類のうち，計算論のレベルおよび表現とアルゴリズムのレベルを意図している．すでに述べたように，複数主体が関与するゲーム的問題状況において，計算論のレベルでは複数主体の「計算」の相互作用を記述する必要があるため，まずは何らかの理論的に扱いやすいアルゴリズムを仮定し，協調問題が解決できる条件を表現とアルゴリズムのレベルにおいて示すことが重要だと考えられる．本学位論文では，理論的な観点から，極端に情報の限定された問題状況として，各主体が自分のアクションとそれに対するフィードバックしかえられない状況を扱う．加えて，こうした不確実性に対処しうるアルゴリズムとして，限定された情報のもと自分に関する情報のみを用いる学習機構である強化学習を用いる．強化学習は報酬信号（利得）をフィードバックとする単純な学習機構でありながら，環境が定常などの条件下では報酬を最大化する最適な行動確率を学習できることが知られている．

学習主体による相互協調の計算論的理解にむけて，本学位論文では，強

[3]その理由としては，各主体が必要な情報を処理しきれない，情報を入手できない，情報を秘匿されている，あるいは情報を信頼できないなどが考えられる．例えば，国家機密，企業秘密のように，相手に情報を開示しないことが今後の展開を有利に進めることに繋がる．

[4]厳密には，「部分的な情報しかえられていない」ことを知っている（全体像だと思い込んでいない）という場合もありうる．本論文で扱うプレイヤは単純な推論しか行わないため，「全体像だと思い込んでいる」と表現できるだろう．

化学習を行う主体が相互協調を実現できる条件を明らかにすることを目的とする．強化学習は期待利得の最大化を目的とする学習機構であるため，強化学習同士のゲームの結果は強化学習のパラメータと利得行列の数値に依存することが予想される．そこで，より具体的には，相互協調の条件として，(a.1) 強化学習のパラメータに関する条件，(a.2) 囚人のジレンマの利得行列に関する条件を明らかにする．また，強化学習は学習の結果として長期的な履歴を条件とした確率的な行動パタンを獲得する．そこで，この確率的な行動パタンを分析することを通して，(a.3) 強化学習が相互協調を実現するときの行動原理を明らかにする．最後に，学習主体による相互協調の計算論的理解にむけて，(b) 以上の知見を先行研究の知見と比較することで，強化学習主体が相互協調を実現する際の鍵となる性質を論じる．

## 1.4 本論文の構成

本論文の構成は以下である．まず，2 章では，相互協調の可能性を理論的に示した先行研究を紹介する．このとき，先行研究の知見から，「長期利益」，「学習」，「不確実性」，「返報性」といった要因が鍵となることを導く．これらの要因のうち，長期利益，学習，不確実性は強化学習において自然に扱える．3 章では，強化学習を戦略とするプレイヤの囚人のジレンマを扱い，(a.1) および (a.3) に対応する分析を行う．4 章では，近似モデルを用いて強化学習戦略の囚人のジレンマを扱い，(a.2) に対応する分析を行う．5 章では，強化学習を含むより一般的な戦略クラスの分析に向けて，いくつかの分析と今後の展開を述べる．6 章では，(a.1)–(a.3) の知見に基づき，(b) に対応する考察を行う．

# 2　先行研究：相互協調の可能性

## 2.1　本章の目的

囚人のジレンマを題材として協調問題に取り組んだ研究を協調問題の解決可能性という観点からまとめ，相互協調を実現可能とする主要因およびその論理を押さえる．本章ではまず囚人のジレンマおよび繰り返し囚人のジレンマの数理的記述を与える．後続の節ではこの数理的記述を使って，先行研究の主要な知見および論理を詳しく紹介する．これらの知見および論理から，協調問題の解決あるいは相互協調の実現に関して，「長期利益」，「学習」，「不確実性」，「返報性」といった要因が鍵となることが予想される．次章ではこれらの要因をみたす戦略として，強化学習戦略を分析する．

## 2.2　囚人のジレンマ

囚人のジレンマでは，$N = 2$ 人のプレイヤは $M = 2$ つの行動，協調（Cooperate: C）または裏切（Defect: D），のうちいずれかを選択し，両プレイヤの行動（行動ペア）に対して各プレイヤの利得（利得ペア）が定められる．プレイヤ $i \in \{1, 2\}$ の行動を

$$x_i = \begin{cases} \mathrm{C} & \text{if Cooperate} \\ \mathrm{D} & \text{if Defect} \end{cases}$$

と記す．両プレイヤの行動ペアは

$$(x_1, x_2) \in \{\mathrm{C}, \mathrm{D}\} \times \{\mathrm{C}, \mathrm{D}\}$$

のいずれかとなる．利得関数 $f_i$ は行動ペア $(x_1, x_2)$ に対してプレイヤ $i$ の利得 $r_i$ を与える．プレイヤ $i$ の利得関数を

$$r_i = f_i(x_1, x_2)$$

と記す．行動ペア $(x_1, x_2) = (\mathrm{X}, \mathrm{Y})$ を XY と略記する．囚人のジレンマの利得行列（表 2.1）は $f_1(\mathrm{DC}) > f_1(\mathrm{CC}) > f_1(\mathrm{DD}) > f_1(\mathrm{CD})$ をみたす（プレイヤ 2 の場合は DC と CD を交換したもの）．

表 2.1: 囚人のジレンマの利得行列（プレイヤ 1）

| | | Cooperate ($x_2 = \mathrm{C}$) | Defect ($x_2 = \mathrm{D}$) |
|---|---|---|---|
| Cooperate | ($x_1 = \mathrm{C}$) | $f_1(\mathrm{CC})$ | $f_1(\mathrm{CD})$ |
| Defect | ($x_1 = \mathrm{D}$) | $f_1(\mathrm{DC})$ | $f_1(\mathrm{DD})$ |

ゲーム理論では，具体的な行動 $\{\mathrm{C}, \mathrm{D}\}$ を純粋戦略という．他方，純粋戦略を確率的に選ぶ戦略を混合戦略という．混合戦略は純粋戦略の集合上の確率分布 $P_i(x_i \in \{\mathrm{C}, \mathrm{D}\})$ で与えられる．ただし，$\sum_{x_i} P_i(x_i) = 1$ かつ $P_i(x_i) \geq 0$ をみたす．混合戦略の利得関数は，戦略ペア $(P_1, P_2)$ に対して，その期待値

$$f_i'(P_1, P_2) = \sum_{x_1} \sum_{x_2} P_1(x_1)\, P_2(x_2)\, f_i(x_1, x_2)$$

と定義される．混合戦略の特殊型として，純粋戦略は $P_i(\mathrm{C}) = 1$ あるいは $P_i(\mathrm{D}) = 1$ に対応する．そこで，混乱のない場合，$f_i$ と $f_i'$ を区別せず $f_i$ で表記する．

ゲーム理論では，独立して戦略を変えることによって，どのプレイヤもより高い利得をえることができないとき，その戦略の組み合わせを Nash 均衡という．Nash 均衡はゲームの解を定義する概念であり，以下のように定義される．ある利得関数 $g_i$ に関して[1] ある戦略の組み合わせ $(s_i^*, s_{-i}^*)$ が Nash 均衡であるならば，すべてのプレイヤ $i$ とすべての戦略 $s_i \in \mathcal{S}_i$ に対して，$g_i(s_i^*, s_{-i}^*) \geq g_i(s_i, s_{-i}^*)$ をみたす．

完全合理的プレイヤを考える場合，各プレイヤは，相手プレイヤの選択が C か D かに依らず，いつでも D を選ぶことが好ましい．混合戦略を考える場合でも，相手プレイヤの C をだす確率に依らず，確率 1 で D を選ぶことが好ましい．したがって，両プレイヤの合計利得を最大にす

[1] プレイヤ集合 $\mathcal{N} = \{1, \ldots, N\}$ に対して，$\mathcal{N} \setminus \{i\}$ を $-i$ と記す．ある戦略の組み合わせ $s = (s_1, s_2, \ldots, s_N)$ に関して，戦略 $s_i$ を除く他のすべての戦略の集合を $s_{-i} = \{s_j : j \neq i\}$ と表記する．利得関数 $g_i$ は添字集合（indexed set）に数値を割り当てる．添字集合とは, 要素 $s_i$ の添字 $i$ で位置が定められている集合で $s = (s_1, s_2, \ldots, s_N) = (s_2, \ldots, s_N, s_1)$ は等価とみなされる．一般的に $s = (s_i, s_{-i}) = (s_j, s_{-j})$ をみたす．

る選択は相互協調 CC であるが，完全合理的プレイヤの推論は相互裏切 DD をもたらす．囚人のジレンマの Nash 均衡は DD のみである．このように囚人のジレンマでは，相手プレイヤが D を選択するとわかっている場合でも，独立して行動を変えることによって利益を高めることができない．また別の概念として，あるプレイヤの効用を低くすることなしに別のプレイヤの効用を高くできないとき，その状態を Pareto 効率的であるという．囚人のジレンマの Pareto 集合は CC，CD，DC である．このように囚人のジレンマでは，個人の合理性（Nash 均衡）と集団の合理性（Pareto 効率性）が両立しない．

## 2.3 繰り返し囚人のジレンマ

繰り返し囚人のジレンマ（IPD）は囚人のジレンマを同じ相手プレイヤと繰り返し行うゲームである．ある時点 $t$ のプレイヤ $i \in \{1,2\}$ の行動を $x_{i,t} \in \{\mathrm{C},\mathrm{D}\}$ と記し，行動ペアを $x_t = (x_{1,t}, x_{2,t})$ と記す．各時点では，両プレイヤの行動ペアは $\mathcal{M} = \{\mathrm{C},\mathrm{D}\}^N = \{\mathrm{CC},\mathrm{CD},\mathrm{DC},\mathrm{DD}\}$ のいずれかとなり，ある時点 $t$ までの両プレイヤの行動ペアの履歴は

$$X_t = (x_{t-1},\ x_{t-2}, \ldots,\ x_1) \in \mathcal{M}^{t-1}$$

となる．ある時点 $t$ でプレイヤ $i$ が受けとる利得は $r_{i,t} = f_i(x_t)$ で与えられる．あるゲームが繰り返し囚人のジレンマであるには，

$$\begin{aligned} &f_1(\mathrm{DC}) > f_1(\mathrm{CC}) > f_1(\mathrm{DD}) > f_1(\mathrm{CD}) \\ &\quad \text{and} \quad 2\, f_1(\mathrm{CC}) > f_1(\mathrm{CD}) + f_1(\mathrm{DC}) \end{aligned}$$

をみたす必要がある（プレイヤ 2 の場合は DC と CD を交換したもの）．ここで，第 2 条件は相互協調 CC が集団的にもっとも望ましい状態であることを保証する．

繰り返し囚人のジレンマは行動計画すなわち「何度も繰り返される囚人のジレンマをどうプレイするか」を戦略（選択肢）とするゲームである．$T$ 回繰り返し囚人のジレンマにおいて，将来の $T$ 回のプレイに関するプレイヤ $i$ の行動計画 $s_i = (x_{i,T},\ x_{i,T-1}, \ldots,\ x_{i,1}) \in \{\mathrm{C},\mathrm{D}\}^T$ を純粋戦略という．純粋戦略は各個体の行動履歴と一対一対応する．他方，混合戦略は純粋戦略の集合上の確率分布 $P_i(s_i \in \{\mathrm{C},\mathrm{D}\}^T)$ で与えられる．ただし，$\sum_{s_i} P_i(s_i) = 1$ かつ $P_i(s_i) \geq 0$ をみたす．

繰り返しゲームにおいても，両プレイヤの戦略（戦略ペア）に対して，各プレイヤへの利得（利得ペア）が定められる．繰り返し囚人のジレンマでは，各プレイヤの受けとる利得は各時点での囚人のジレンマの利得の合計となる．すなわち，戦略ペア $(s_1, s_2)$ に対するプレイヤ $i \in \{1, 2\}$ の平均利得は，その戦略が定める行動ペア $(x_{1,t}, x_{2,t})$ を用いて，

$$F_i(s_1, s_2) = (1/T) \sum_{t=1}^{T} f_i(x_{1,t}, x_{2,t})$$

と定義できる．また，混合戦略の利得関数は

$$F'_i(P_1, P_2) = \sum_{s_1} \sum_{s_2} P_1(s_1)\, P_2(s_2)\, F_i(s_1, s_2)$$

と定義できる．混乱のない限り，$F_i$ と $F'_i$ を区別せず $F_i$ で表記する．$T$ 回繰り返しゲームにおいて，最後の時点を含めた履歴 $X_{T+1}$ は純粋戦略ペア $s = (s_1, s_2)$ と一対一対応する．そこで，利得関数の表記を $F_i(X_{T+1}) = F_i(s) = F_i(s_1, s_2)$ と拡張する．繰り返しゲームの Nash 均衡は $F_i$ に関して定義される．

**行動戦略**　繰り返しゲームの戦略は任意の仕方で記述できるが，過去のプレイの結果（情報集合）を条件として次の行動を決める戦略を行動戦略という．行動戦略は各時点 $t = 1, \ldots, T$ の各履歴（情報集合）に対して $x_{i,t}$ を選択する確率を割り当てる関数 $P_i(x_{i,t}|X_t) \in [0, 1]$ として定義できる．完全記憶の場合，行動戦略にはそれと等価な混合戦略が存在し，行動戦略の利得関数は混合戦略に変換してえられる．

**マルコフ戦略**　過去 $K$ 時点前までの履歴に依存して次の行動 $x_{i,t}$ を選択する確率を割り当てる行動戦略を $K$ 次マルコフ戦略（$K$ 次戦略）という．過去 $K$ 時点前までの両プレイヤの行動ペアの履歴を

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

と記すとき，$K$ 次マルコフ戦略 $P_i$ はマルコフ性 $P_i(x_{i,t}|X_t^K) = P_i(x_{i,t}|X_t^\infty)$ をみたす．混乱のない場合，$X_t$ と $X_t^K$ を区別せず $X_t$ で表記する．

マルコフ戦略は状態空間 $X \in \mathcal{M}^K$ 上の遷移確率行列 $Q$ として表現でき，初期確率分布 $\pi_0$ からの時間発展は $\pi_{t+1} = Q\,\pi_t$ となる．ここで，遷

移確率行列 $Q$ は戦略ペア $(P_1, P_2)$ の関数である．時間の極限 $T \to \infty$ でえられる確率分布 $\pi_\infty$ を定常確率分布といい，$\pi_\infty = Q\,\pi_\infty$ をみたす．

マルコフ戦略の利得関数はマルコフ過程の確率分布 $\pi_T$ を用いて，

$$F_i''(P_1, P_2) = \sum_{X \in \mathcal{M}^K} \pi_T(X)\, F_i(X)$$

と定義される．ここで，$\pi_T(X)$ は履歴 $X$ が実現される確率を表す．混乱のない限り，$F_i$，$F_i'$，$F_i''$ を区別せず $F_i$ で表記する．

### 2.3.1 有限回繰り返し囚人のジレンマ（後方帰納）

有限回 IPD では相互裏切が唯一の Nash 均衡であることが知られている．繰り返し回数を $T$ とする．プレイヤ $i$ の目的は将来の $T$ 回のプレイで最適な戦略 $s_i = (x_{i,T}, x_{i,T-1}, \ldots, x_{i,1})$ を見つけることである．この問題は後方帰納（動的計画法）により解ける．まず最後の時点 $T$ での行動を最適化するが，これは最後のプレイなので将来報復される可能性はなく，したがって標準型の囚人のジレンマと同じで裏切 D が最適行動である．次に時点 $T-1$ での行動を最適化するが，現時点で何をしようが次時点で相互裏切 DD となるならば，裏切 D が最適行動である．再帰的に推論することで，最初の時点 $t = 1$ から裏切 D を選択する戦略が最適戦略だとわかる．この戦略を ALLD を呼ぶ．ALLD は有限回 IPD のだたひとつの Nash 均衡である．

### 2.3.2 無限回繰り返し囚人のジレンマ（素朴定理）

無限回 IPD では相互協調が無限に多く存在する Nash 均衡のひとつになりうることが知られている [23]．無限回 IPD は最後の時点が存在せず，将来報復される可能性がいつでも現在の意思決定に影を落とすため，将来の利得をどう扱うかが重要である．ここでは，素朴定理により与えられる相互協調 CC が Nash 均衡になる条件とその論理を整理する．

合理的プレイヤ $i$ が意思決定を行うときは $F_i$ を最大化するよう戦略（あるいは行動計画）を選択するが，そのためには各戦略の現在価値を見積もる必要がある．現在価値は将来価値に関して時間に依存した割引 $\eta^t$ $(0 \le \eta \le 1)$ を適用するものが一般的である [2]．すなわち，戦略ペア

[2]割引率を $\eta^{t-1}$ として定義する場合もある．現在価値を計算する時点を $t = 0$ とするか，$t = 1$ とするかの違いである．素朴定理の結論には影響しない．

$s = (s_1, s_2)$ に対するプレイヤ $i$ の戦略 $s_i$ の現在価値は

$$\hat{F}_i(s_1, s_2) = \sum_{t=1}^{T} \eta^t f_i(x_{1,t}, x_{2,t})$$

と定義できる．時間の極限 $T \to \infty$ では次式となる [3]．

$$\hat{F}_i(s_1, s_2) = \sum_{t=1}^{\infty} \eta^t f_i(x_{1,t}, x_{2,t}) = \frac{\eta}{1-\eta} f_i(x_{1,t}, x_{2,t})$$

いつでも C をだす戦略を ALLC と呼ぶ．両プレイヤが ALLC を採用する場合，D をだすことで利益を高められるゆえに，ALLC は均衡ではない．相互協調を維持するには，相手が短期的な利益に惑わされないよう，永遠の報復による長期的な損失の可能性を知らしめる必要がある．この戦略を GRIM またはトリガ戦略という．GRIM は相手が一度でも D をだすまでは C だが，もし相手が一度でも D をだせば二度と C を選ばない．GRIM が相互協調を実現することをみるために，いまある時点 $t$ まで CC のみが出現していたとしよう．問題は，次の時点で相手 GRIM が C をだすとしたら，自分は C か D かどちらを選ぶかである．このまま C を選択するという戦略の現在価値は $\eta f_i(\mathrm{CC}) + \frac{\eta^2}{1-\eta} f_i(\mathrm{CC})$ である．他方，次の時点で D を選択するという戦略の現在価値は，永遠の報復を受けることになり，$\eta f_i(\mathrm{DC}) + \frac{\eta^2}{1-\eta} f_i(\mathrm{DD})$ である．協調し続けるべき条件は

$$\begin{aligned} & f_i(\mathrm{CC}) + \frac{\eta}{1-\eta} f_i(\mathrm{CC}) > f_i(\mathrm{DC}) + \frac{\eta}{1-\eta} f_i(\mathrm{DD}) \\ \iff & \eta > \frac{f_i(\mathrm{DC}) - f_i(\mathrm{CC})}{f_i(\mathrm{DC}) - f_i(\mathrm{DD})} \end{aligned}$$

となる．この意思決定問題にはすべての時点で直面するので，ある時点でこの条件をみたせば，すべての時点でこの条件をみたす．囚人のジレンマでは $f_i(\mathrm{DC}) > f_i(\mathrm{CC}) > f_i(\mathrm{DD})$ より，右辺は $(0, 1)$ の範囲に収まる．したがって，相手の戦略が GRIM だと既知で，$\eta < 1$ が十分大きいとき，相互協調は Nash 均衡となる．

## 2.4 進化ゲーム理論

自然選択はある種の「合理性」をもち，ある環境のもとでより高い利益をえる個体の頻度を増やす．進化ゲーム理論は戦略の集団を用意し，戦略

[3] $\hat{F}_i(s) - \eta \hat{F}(s) = (1 - \eta^T)\, \eta\, f_i(x_{1,t}, x_{2,t})$ より導く．

同士を互いに競わせることで，より高い利益を生みだす戦略を探索するアプローチである．進化ゲーム理論の方法は，遺伝的アルゴリズム，頻度分布のダイナミクス，空間構造を扱うものまで多岐にわたるが [see 42]，その分析と解釈の容易さから１次マルコフ戦略が扱われることが多い．本節では代表的な１次戦略である TFT と WSLS をその行動原理に留意して述べる．どちらもより複雑な推論モデル・行動モデルを単純化したものであるが，それは後続の節で述べる．

Axelrod [3] は，研究者から戦略を公募し，IPD を用いて集められた戦略を合計利得でランクづけした．その結果，TFT（Tit-for-Tat）と呼ばれる，しっぺ返し戦略が高い合計利得をえることが知られている．しっぺ返し戦略 TFT は返報性すなわち「君が協調（裏切）するなら僕も協調（裏切）する」を行動原理とする．しっぺ返し戦略は，協調しながらも侵略や搾取に対抗できる，協調と裏切を兼ね備えた行動戦略である．しかしながら，しっぺ返し戦略はノイズに弱いことが知られており，偶然に裏切が選ばれると相互裏切の循環に陥り，長期的には低い利益しかえられない．そこで，相手も TFT である場合に，搾取の誘因を抑えつつも相互協調へ復帰する機会を与える「寛容性」が提案された [39]．この戦略は GTFT（Generous TFT）と呼ばれる．GTFT はノイズのある状況でも高い合計利得をえることが示されている [45].

進化ゲーム理論の文脈で GTFT を上回る戦略として WSLS（Win-Stay, Lose-Shift）が提案された [46]．WSLS は「負け（痛み）ならば行動を切り換え，勝ち（喜び）ならば前回と同じ行動を続ける」という「痛みを避け，喜びを求める」行動原理に従う [4]．WSLS は PAVLOV [32] と呼ばれる学習戦略を単純化したものであるが，直前の結果のみに応答して行動する．WSLS 同士では相互裏切から相互協調へ次点で復帰できるという性質をもち，ノイズのある状況でも高い合計利得をえるが，ALLD に対しては一方的な搾取を許容する．

数多くの研究から，IPD では戦略の頻度分布が複雑なダイナミクスを示すことが知られている．例えば，戦略 X を戦略 Y が侵略することを X $\to$ Y と記すとき，単純な三角関係として $\cdots \to$ ALLC $\to$ ALLD $\to$ TFT $\to$ ALLC $\to \cdots$ となる（侵略の可能性は変異パラメータに依存する）[5, 42]．また，近年では Zero-Determinant (ZD) 戦略と呼ばれる，２人ゲームにおいて，相手の利益の上限を制御できる戦略が調べられてい

[4] 囚人のジレンマの利得行列は $f(\mathrm{DC}) > f(\mathrm{CC}) > f(\mathrm{DD}) > f(\mathrm{CD})$ をみたすが，WSLS は平均利得を基準に DD と CD を負け，DC と CC を勝ちとみなす．

る [47]．しかし，さまざまな戦略がランダムに対戦しその合計利得を競う進化シミュレーションでは，ZD 戦略は必ずしも勝ち残れるわけではないことが示されている [57]．

## 2.5 メタゲーム理論

TFT はメタゲーム理論 [Howard 27] と呼ばれる自他入れ子型の推論から着想をえて１次戦略として提案された [11]．留意点として，メタゲーム理論は標準ゲームに関する理論である．本節では，原典での TFT の理論的背景を紹介し，相互協調の論理を求める．本節では $x = \sigma_i(y)$ でプレイヤ $j \neq i$ の行動 $y$ に対するプレイヤ $i$ の応答 $x$ を表す（$x$，$y$ はともに戦略）．ただし，$\sigma_i(\varnothing)$ は標準型ゲームで選択する戦略を表す．

標準型囚人のジレンマを基本ゲーム ∋ と呼ぶ．純粋戦略は

$$s_i^0 := \sigma_i(\varnothing)$$

である．これをメタゲーム理論では基本戦略という．基本戦略の集合は $s_i^0 \in \mathcal{S} = \{\mathrm{C}, \mathrm{D}\}$ である．

$j$-メタゲーム $j$∋ とは $i$ の基本戦略に対する $j$ の１次メタ戦略のゲームである．１次メタ戦略とは $i$ の基本戦略 $s_i \in \mathcal{S}$ に対して自分のとりうる反応 $s_j \in \mathcal{S}$ の網羅的記述である．１次メタ戦略を

$$s_j^1 := \sigma_j(\mathrm{C})/\sigma_j(\mathrm{D})$$

と記す．１次メタ戦略の集合は $s_j^1 \in \mathcal{S}^1 = \{\mathrm{C/C}, \mathrm{C/D}, \mathrm{D/C}, \mathrm{D/D}\}$ であり，$|\mathcal{S} \times \mathcal{S}| = 2^2 = 4$ 通りある．

$i$-$j$-メタゲーム $ij$∋ とは，同じように，$j$ の１次メタ戦略に対する $i$ の２次メタ戦略のゲームである．２次メタ戦略を

$$s_i^2 := \sigma_i(\mathrm{C/C})/\sigma_i(\mathrm{C/D})/\sigma_i(\mathrm{D/C})/\sigma_i(\mathrm{D/D})$$

と記す．２次メタ戦略は $|\mathcal{S}^1 \times \mathcal{S}| = 4^2 = 16$ 通りある．

戦略ペアが決まれば，基本ゲーム ∋ での帰結が導けて，そこから利得が定まる．計算の例として，12∋ を考えよう．12∋ はプレイヤ 1 の推論である．このとき，戦略ペア $s_1^2 = \mathrm{C/C/D/D}$ と $s_2^1 = \mathrm{C/D}$ の帰結を導く．プレイヤ 1 の反応はプレイヤ 2 の戦略 C/D より，$\mathrm{C} = \sigma_1(\mathrm{C/D})$ となる（$s_1^2$ の 2 番目の要素）．これを受けて，プレイヤ 2 の反応は $\mathrm{C} = \sigma_2(\mathrm{C})$

となる（$s_2^1$ の 1 番目の要素）．したがって，基本ゲームの帰結は相互協調 CC であり，利得は $f_1(\mathrm{CC}), f_2(\mathrm{CC})$ となる．この計算過程は

$$(\mathrm{C/C/D/D, C/D}) \to (\mathrm{C, C/D}) \to (\mathrm{C, C})$$

のように，戦略ペアの交互反復的な簡約として書ける．

この計算をすべての 2 次メタ戦略と 1 次メタ戦略の組み合わせ（$16 \times 4$ 通り）に対して行えば，12∋ の利得行列を構築でき，この利得行列を標準型ゲームと同様に分析することで，メタゲーム 12∋ の Nash 均衡を調べられる．その結果，12∋ には Nash 均衡が 3 つ存在することがわかる．ひとつは例に用いた $(s_1^2, s_2^1) = (\mathrm{C/C/D/D, C/D})$ であり，他には，$(\mathrm{D/C/D/D, C/D})$，および $(\mathrm{D/D/D/D, D/D})$ がある．最初の 2 つは相互協調となり，最後の 1 つは相互裏切となる．最初の C/D を Howard [27] は tit-for-tat と解釈した [5]．最後の D/D は ALLD と解釈できる．

メタゲームにおける上記 3 つの均衡を標準型ゲームまで簡約してえられる帰結は CC と DD であり，$f_i(\mathrm{CC}) > f_i(\mathrm{DD})$ となる．それゆえに，もし完全情報のもと両プレイヤが 2 次メタ戦略まで推論したとすれば，標準型ゲームの帰結は相互協調となる．メタゲーム理論は計算論のレベルの説明すなわち「自他入れ子型の推論に基づいた標準ゲームにおける利得の最大化」を与えるように思われる．このように，メタゲーム理論は利得行列や他者の戦略に関する情報を前提としている．

## 2.6 学習主体のゲーム

個体が学習能力を有し適応的に振る舞えれば，ゲームの帰結はどう変わるだろうか．von Neumann and Morgenstern [63] は，ゲームを複数の経済主体が各々の目的関数を独立かつ同時に最大化する問題と表現している．この表現を借りれば，学習主体のゲームは，複数の学習主体が同じゲームをプレイし，その試行錯誤の経験から行動を変化させ，各々の目的関数を独立かつ同時に最大化する問題といえる．とくに繰り返し囚人のジレンマでは，どんな相手の戦略にも最適となる唯一の戦略は存在しないことが知られている [4]．換言すれば，最適な戦略は必ず相手の戦略に依存して決まり，一方の戦略を固定することなしに他方の戦略を最

[5] Axelrod [3] トーナメントに TFT を提出した Anatol Rapoport はメタゲーム理論の支持者であった [11].

適化できない [4]．これらは相互に学習を行う主体同士のゲームがもつ根本的な複雑性を表現している．

学習主体のゲームで初期のものは Brown [12] による信念学習（belief learning）である．信念学習は各行動の累積価値に比例した確率で行動を決め，各時点で相手の行動に対する最適応答を強化する．信念学習は完全情報のもとでの学習であり，学習の結果は完全合理的プレイヤの１回ゲームと同じく相互裏切となる．

別の系統として，Bush and Mosteller [13] による要求学習（learning with aspiration）[6] がある．要求学習では要求レベルという変数をもち，外部報酬は要求レベルから相対的に評価され，行動確率の更新に使われる [7]．要求学習を用いた IPD では，学習の結果，初期条件によっては相互協調へ収束しうる [32, 34, 48]．PAVLOV は要求レベルを $f_1(\mathrm{CC})$ と $f_1(\mathrm{DD})$ の間に設定し，要求レベルを超えるか否かで喜びと痛みを切り分ける [32]．WSLS は PAVLOV のこの点を継承した１次戦略である [46]．PAVLOV は TFT や PAVLOV との対戦において高い確率で相互協調を学習できる [32]．また，要求レベルを動的に変更するモデルは ALLD に対して搾取させず相互裏切へ収束できる [36]．

強化学習（reinforcement learning）[Sutton and Barto 58] は，行動に対して報酬を受けとり，各行動を累積報酬に比例した確率で選択する．強化学習は工学的問題の解決や生物の行動の説明など数多くの実績をもち，理論的基礎の確立された学習ルールのひとつである．強化学習は将来の期待利益を最大化するという目的のもと，試行錯誤の経験から行動確率を最適化できることが知られている [58]．しかしながら，強化学習を用いた IPD 研究は少なからずあるが，学習の結果として高い確率で相互協調を実現できたという報告は少ない．Sandholm and Crites [50] は Q 学習とニューラルネットの対戦を分析し，ほぼランダムか相互裏切という結果をえている．Masuda and Nakamura [36] は $\varepsilon$-greedy TD 学習を分析し，前回の自分の行動を状態として学習する場合は相互裏切となるが，他方，前回の自分と相手の行動ペアを状態として学習する場合は相互協調の確率が高まることを示した．また類似した結果として，Banerjee and

[6] 本論文では Sutton and Barto [58] などによる強化学習と区別するため「要求学習」と呼ぶが，該当の分野では「強化学習」と呼ばれる．

[7] Bush and Mosteller [13] の定式化は (a) 確率と報酬を足し合わせたり，(b) 行動確率の変数が負の値をとらないよう最大・最小を調整するなど奇妙な点がある．また，Roth and Erev [49] の定式化も類似した問題をもつ．要求学習を精緻化する取り組み [17, 22] があるが，こうした問題は Sutton and Barto [58] による強化学習にはない．

Sen [6] は $\varepsilon$-greedy Q 学習（拡張あり）を分析し，自分の行動だけの場合は相互裏切だが，相手の行動との条件付期待報酬に比例した行動確率の場合，相互協調の確率が高まることを示した．また，期待利得の微分情報を使用し，強化学習を勾配降下法で近似的に解く IGA/GIGA アルゴリズム [55, 66] や WoLF 原理 [10] と呼ばれるアプローチでは，Singh et al. [55] は相互裏切へ収束することを示した．

以上の知見から，IPD ゲームの結果は学習理論によって大きく変わりうることがわかる．信念学習は完全情報を前提とするが，要求学習と強化学習は必ずしも完全情報を要求せず，どこまでの情報を主体が利用できるかは研究によって異なる．要求学習は，相互協調の可能性が報告されているが，「痛みを避け，喜びを求める」という各時点の行動変化（表現とアルゴリズムのレベル）のみを記述しており，必ずしも学習の目的（計算論のレベル）が定義されていない．他方，強化学習は自己利益（累積報酬）の最大化という目的をもつ．強化学習を用いた研究では，情報の可視性の影響が分析されているが，相手の行動と関連づけた学習の必要性を示唆していており，相手の行動と自分の利得を利用可能とする場合のみ相互協調が報告されている．

## 2.7 ゲーム構造の改変

協調問題を，外部から介入し問題そのものを改変することで解消しようという試みがある．介入はより協調的な別のゲームに改変することを意図する．Cabrera and Cabrera [15] は (a) 利得行列の再編成，(b) 自己効力感の向上，(c) 集団認識と個人責任の向上をあげている．企業経営では，部下の協調問題を解決したい，部下の連携を強化したいという要望がある．利得行列の再編成は，上司の立場から，行動に応じた特別な報酬を部下に与え，部下たちの利害関係を改変しようとする．自己効力感の向上は，部下に連携と成功の体験を教えることで，個人の協調の水準を引き上げようとする．最後に，集団認識と個人責任の向上は，グループでの問題意識の共有，グループにおける個人の役割の共有を促すことで，集団的活動への参入意欲を高めようとする．

利得行列の再編成の一例として，囚人のジレンマにおける利害の公平性（fairness）を考えてみよう．外部から利害を調整し，仮に囚人のジレンマで両プレイヤが公平な利得（平均利得）$g(\mathrm{XY}) := [f_1(\mathrm{XY}) + f_2(\mathrm{XY})]/2$ をえるとしよう．このとき，本来の囚人のジレンマでは $f_1(\mathrm{DC}) > f_1(\mathrm{CC}) >$

$f_1(\mathrm{DD}) > f_1(\mathrm{CD})$ であるが，この「公平な囚人のジレンマ」では $g(\mathrm{CC}) > g(\mathrm{DC}) = g(\mathrm{CD}) > g(\mathrm{DD})$ と改変され[8]，両プレイヤにとって協調 C が支配戦略となる．また別の例として，選択肢を制限し，TFT と ALLD のみを選択肢（戦略）とする無限回 IPD は，各戦略ペアの期待利得を計算すると標準型の鹿狩りゲームと一致することがわかり，相互協調を実現できる戦略ペア (TFT, TFT) が支配的となる [56].

以上の知見は，協調問題の解決という点では現実的かつ実践的示唆を与えるものの，協調問題を理論的により協調しやすいとされる別の問題に掏り替え，協調問題自体を直接的に解消していない．また，掏り替えた別の問題を解決する必要が生じる．

## 2.8 ヒトの限定合理的な意思決定

多くの経験的な研究から，ヒトの意思決定は必ずしもゲーム理論の想定する完全合理性を満たさないことが指摘されている [16, 17, 30, 49]. Kahneman and Tversky [30] は損益が確率的に与えられる不確実な意思決定状況を心理実験で検証し，古典経済学の仮定する期待効用の最大化とは異なり，人は利益を期待するときと損益を期待するときで異なる行動選択をすることを示した.

心理学・行動経済学の実験からは，繰り返しゲームにおいて人が暗黙にもつゲームの継続性に対する不確実性やゲームの利得や対戦相手に関する情報の有無が意思決定に影響を与えることが示されてきた．繰り返し囚人のジレンマに関する理論的および実験的研究では，期待されるゲームの繰り返し回数が多いほど，協調の確率が高まることが示唆されている [7]．プレイヤがゲームの繰り返し回数を知っている場合，理論的には完全合理的なプレイヤは初回から裏切るとされる．しかし，有限回繰り返しゲームであっても，人は序盤は協調し，終盤に近づくほど裏切ることが経験的に示されている [2, 19, 53]．こうした人の行動を説明するひとつの要因は相手の利得行列に関する不確実性であると考えられる．相手の利得行列に関して不確実性が高いときには，プレイヤの選択が協調 C から裏切 D へと，ゲームの回数に応じて変化することが理論的に示されおり [33]，また実験的にも確認されている [28, 29].

[8] 利得がプレイヤ間で対称のときでも，IPD では $2f_1(\mathrm{CC}) > f_1(\mathrm{CD}) + f_1(\mathrm{DC})$ は保証されるが，$f_1(\mathrm{CD}) + f_1(\mathrm{DC}) > 2f_1(\mathrm{DD})$ は保証されていない.

以上の知見を踏まえると，合理的な意思決定により生じる個人と集団の間のジレンマが，ゲームに関する部分的な情報とそれを補う推論に基づく限定合理的な意思決定では生じない可能性がある．

## 2.9 まとめ

協調問題に関する先行研究の知見および論理を以下にまとめる．

**完全合理性と囚人のジレンマ** 相互裏切は唯一の Nash 均衡である．

**完全合理性と有限回 IPD** 相互裏切 ALLD は唯一の Nash 均衡である．

**完全合理性と無限回 IPD** 相互協調は無限に存在する Nash 均衡のひとつである．相互協調には，相手の裏切を牽制しながら，十分に長期的な利益を考慮する必要がある．素朴定理は GRIM の原点である．

**進化ゲーム理論** TFT や WSLS という戦略は同胞に対して相互協調を実現でき，高い期待利得をもたらす．TFT は「君がそうするなら僕もそうする」，WSLS は「痛みを避け，喜びを求める」という行動原理に従う．

**メタゲーム理論** 完全合理的な自他入れ子型の推論によって，相互協調を可能とするメタ戦略が存在する．その戦略が TFT の原点である．

**学習主体のゲーム** 信念学習（完全情報）では相互裏切となる．要求学習（不完備）のひとつ PAVLOV は WSLS の原点である．強化学習（不完備）は相手の行動と関連づけて学習すれば相互協調が可能である．

**ヒトの限定合理性** 理論に反してヒトは有限回 IPD でも序盤は協調する．繰り返し回数や利得行列が不確実の場合，協調が促進される．

以上のまとめから，協調問題の解決あるいは相互協調の実現に関して，「長期利益」，「学習」，「不確実性」，「返報性」といった要因が鍵となることが予想される．これらのうち「長期利益」や「学習」は，現実においては認知主体の能力を特徴づけるものであるが，ゲーム理論では戦略の性質を特徴づけるものといえる．本論文では，囚人のジレンマの利害構造

を基礎として，限定合理性を採り入れたことによるジレンマの解消について論じる．具体的には，限定合理的な意思決定の状況として，それぞれのプレイヤが自身への利得が相手のプレイヤの選択に依存することを知らない状況を考える．この状況では，各プレイヤの選択に対して複数の利得があり，一種の不確実性の下での意思決定の問題として捉えられる．本論文では，こうした不確実性のある問題状況において，不確実性を克服する手段としての個体の学習を扱い，それにより個人と集団のジレンマが解消されるかを議論する．

# 3　強化学習戦略による協調問題の解決

## 3.1　本章の目的

繰り返し囚人のジレンマの各ゲームでは，2 人のプレイヤの選択の組み合わせにより各プレイヤへの利得が決まる．また各プレイヤは過去にえた利得をもとに次なる行動の選択を行う．本研究では，過去の行動選択とそれに対する利得から次の意思決定を確率的に行うモデルとして，強化学習を用いる．利得がプレイヤの戦略に応じて変化しない条件の下において，ある種の強化学習は個人の利得を最大化することが知られているため [58]，本研究では不確実な状況において個人の利得を最大化する戦略の近似とみなして，強化学習戦略の分析を行う．この戦略は学習し動的に変化する相手プレイヤの存在を各プレイヤが知らないという限定合理性を反映する．強化学習戦略をとるプレイヤ同士の繰り返しゲームは一種の確率過程とみなせる．本研究ではこの確率過程の長期的な構造の分析により，個人と集団の意思決定における最適性の関係を明らかにする．

## 3.2　定式化

本研究では大別して，相手の行動に応じて次の行動を選択する戦略と，相手の行動によらずに次の行動を選択する戦略を分析する．前者の戦略は 1 回前のゲームの帰結に基づき意思決定するため，これを 1 次戦略と呼ぶ．後者の戦略は過去の自らの行動に対する利得に応じて行動選択を行う強化学習の枠組み [58] に従うため，これを強化学習戦略と呼ぶ．本研究では 1 次戦略と強化学習戦略を併せて調べることで，1 次戦略に関する先行研究の知見を強化学習戦略を理解するために応用し，強化学習戦略のもつ限定合理性の性質を分析する．

### 3.2.1 1次戦略

Nowak [41] は IPD において1回前のゲームの帰結に基づき意思決定する戦略を分析した．ある結果 $y$ が実現したあとで $x$ が生起する条件付確率を $P_i(x|y)$ と記すとき，1次戦略は時点 $t$ のゲームの帰結 $x_t \in \mathcal{M} = \{\mathrm{CC}, \mathrm{CD}, \mathrm{DC}, \mathrm{DD}\}$ に対する応答として次に協調 C をとる確率 $P_i(x_{i,t+1} = \mathrm{C}|x_t)$ により表現できる．1次戦略を含めて本研究で対象とする戦略は，時点 $t$ によらず一定であると仮定するため，以後，表記を単純化して，ある帰結 $y \in \mathcal{M}$ を受けて C をだす確率を $P_i(\mathrm{C}|y)$ と書く．

本研究で比較する7つの1次戦略を表 3.1 に列挙する．先行研究ではこれらの1次戦略が分析されている [5, 41, 45, 46]．経験的な分析から，他のすべての1次戦略に優る万能な戦略は見つかっていないが，しっぺ返し戦略 TFT（Tit-for-Tat）は相対的に高い平均利得をえることが知られている．一部の戦略を除いて，1次戦略は基本的に直前の相手の行動に反応して次の行動を選択する．

ここでは簡潔に各1次戦略について述べる（3.7 節で詳しく分析する）．ALLC はかならず C をとるという戦略である．ALLD はかならず D をとるという戦略である．RAND は常に C と D を等確率（1/2）でとる戦略である．GRIM は，履歴がすべて CC の場合を除いて D をとる戦略である．TFT（Tit-for-Tat）は最初は C をとり，以後は相手の直前の行動を模倣するという戦略である．GTFT（Generous TFT）は TFT とほぼ同じだが，相手が D のときでも一定の確率（1/3）で C をとるという戦略である．WSLS（Win-Stay, Lose-Shift）は「負け」なら行動を切り替え，「勝ち」なら前と同じ行動をとるという戦略である [1]．

### 3.2.2 強化学習戦略

囚人のジレンマでは各プレイヤの利得が相手の選択に依存しており，合理的主体や1次戦略の一部は各プレイヤがその利得行列を既知であることを前提に意思決定を行う．一方，本研究では部分的なゲームに関する情報とそれを補う推論に基づく限定合理性の一種として，各プレイヤがゲームの利得行列の情報をもたずに [2] 行動を選択する場合を考える（図 3.1）．この場合，各プレイヤは自身の選択した行動とそれに対する利得のみを

[1] Nowak and Sigmund [46] では，利得行列の平均値を基準に，平均値以上の利得をもつ DC，CC を勝ち，平均値以下の利得をもつ DD，CD を負けとしている．

[2] ゲーム理論の用語では不完備情報ゲームという．

表 3.1: 1 次戦略をとるプレイヤ 1 が 1 回前のゲームの帰結に応じて次に協調 C を選択する確率 $(P_1(\mathrm{C}|\mathrm{CC}), P_1(\mathrm{C}|\mathrm{CD}), P_1(\mathrm{C}|\mathrm{DC}), P_1(\mathrm{C}|\mathrm{DD}))$

| 戦略名 | 協調 C の選択確率 |
|---|---|
| ALLC | ( 1 , 1 , 1 , 1 ) |
| ALLD | ( 0 , 0 , 0 , 0 ) |
| RAND | ( ½ , ½ , ½ , ½ ) |
| TFT | ( 1 , 0 , 1 , 0 ) |
| GTFT | ( 1 , ⅓ , 1 , ⅓ ) |
| GRIM | ( 1 , 0 , 0 , 0 ) |
| WSLS | ( 1 , 0 , 0 , 1 ) |

知る．つまり，プレイヤ 1 からは 2 種類の行動ペア $(x_{1,t}, x_{2,t}) = (\mathrm{C}, \mathrm{D})$ と (C, C) を区別できず，自分の選んだ行動 C に対して未知なる相手の行動（D または C）に依存した利得 $f_i(\mathrm{CD})$ または $f_i(\mathrm{CC})$ をうけとる．したがって，本研究では過去の自分の選択とそれに対する自分への利得だけの関数として次の時点における選択確率を与える戦略を考える．

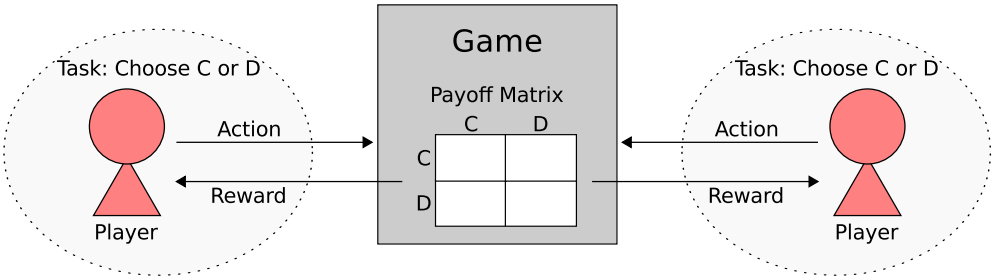



図 3.1: 強化学習戦略をとる主体のゲーム

各プレイヤが確率的に意思決定をする場合，各プレイヤの選択に対する利得は確率的に決まる．こうした不確実な状況の下で，相手やゲームの情報を要求せず，利得という個人的な効用だけに基づいて漸近的により大きな利得をもたらす行動に高い確率を割り当て，期待利得を最大化する戦略のひとつとして強化学習戦略が研究されている [26, 50, 52, 68]．本研究では 1 次戦略に加えて，以下に紹介する強化学習戦略を分析した．これらの各 1 次戦略との対戦において，3.7 節で示すとおり，強化学習戦略は高い期待利得をえられるため，本分析では強化学習戦略を個人の利

得を最大化する戦略として扱う．ただし，本研究は利得が不可知な相手に依存する限定合理的な状況における学習を扱うため，必ずしもこの状況での利得の最大化を意味しない．本研究では「最大化」を，プレイヤ自身の戦略が平均利得を変えないなど，ある十分に制約された状況において期待利得を最大化するという意味で用いる．

強化学習戦略では，ある時点 $t$ 直前までの行動ペアの履歴

$$X_t = (x_{t-1},\ x_{t-2}, \ldots,\ x_1)$$

に対して，行動 $x$ に関する，時間に応じて割引した累積利得

$$R_{i,x}(X_t) = \lim_{K\to\infty} \sum_{k=1}^{K} \alpha_i^k\, \delta(x, x_{i,t-k})\, r_{i,t-k} \tag{3.1}$$

を考える．この定式化は強化学習の分野では逐一訪問モンテカルロ法と呼ばれる [58]．式 (3.1) において，

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

であるため，$R_{i,x}(X_t)$ はプレイヤ $i$ が時点 $t-k$ で行動 $x_{i,t-k}$ を選択した場合に，その行動に対する利得 $r_{i,t-k} = f_i(x_{t-k})$ を時間に依存する割引係数 $\alpha_i^k$ で重み付けたものの総和を表す．累積利得はパラメータ $0 \leq \alpha_i \leq 1$ の関数であり，$\alpha_i$ が大きいほど，過去の利得をより高く評価する．

強化学習戦略は累積利得 $R_{i,x}(X_t)$ と鋭敏性パラメータ $\beta_i \geq 0$ をもつロジスティック関数として以下のように与えられる：

$$P_i(x|X_t) = \frac{\exp[\beta_i\, R_{i,x}(X_t)]}{\sum_y \exp[\beta_i\, R_{i,y}(X_t)]}\ . \tag{3.2}$$

鋭敏性 $\beta_i$ が大きいほど累積利得の高い行動をより高い確率で選択し，$\beta_i = 0$ の場合は累積利得によらず C か D を確率 $1/2$ で選択する．

### 分析対象

強化学習戦略は $\alpha_i$，$\beta_i$，$K$ というパラメータをもつ．このうち本研究ではとくにゲームの結果に強い影響をもつパラメータ $\alpha_i$ に着目した分析を行った．

本研究において主たる分析対象とするパラメータ $\alpha_i$ を「記憶保持率」と呼ぶ．式 (3.1) は漸化式 $R_{i,x}(X_{t+1}) = \alpha_i \left(R_{i,x}(X_t) + \delta(x, x_{i,t})\, r_{i,t}\right)$ として表現できる．したがって，$\alpha_i$ は 1 時点前の利得 $r_{i,t}$ と累積利得の和をどれくらい保持するかを決めるパラメータである．また本稿の定式化では $\alpha_i = 0$ の場合は累積利得は常に 0（記憶なし）となり，C と D を等確率で選択する．$\alpha_i$ が定数である場合，厳密には利得を最大化する行動に確率 1 で収束しない [58]．より厳密な議論には $\alpha_i$ をステップ数の関数とする必要があるが，この議論は本論文の主旨には必要ないため，本研究の分析では記憶保持率 $\alpha_i$ を定数として分析した．

式 (3.1) は十分長い利得の履歴（$K \to \infty$）から累積利得を計算することを意味する．本研究では 3 節で示すように $K$ 次マルコフ過程として分析する．$K$ 次マルコフ過程の状態数は $K$ の指数関数 $4^K$ で増大するため，定常分布の計算資源の限界のため $K = 10$ とし，十分な長さをもつ強化学習の近似とみなした．この近似の妥当性については著者らの研究 [26, 68] で議論している．履歴長 $K$ が十分に大きい限り，実質的に結果に影響しないため，本論文では分析の対象としない．

$\beta$ の影響を分析した著者らの研究 [68] では，相互協調の発生に関して定性的に本稿と同じ結論がえられている．このため，本分析では $\beta = 1$ に固定し，このパラメータに関する分析を行わない．

## 3.3 分析方法

### 3.3.1 有限状態マルコフ過程

繰り返し囚人のジレンマの各時点は 4 つの状態 CC，CD，DC，DD をもち，その状態の系列に対して次の状態への遷移確率が一意に定まる．このような確率過程はマルコフ過程と呼ばれ，とくに $K$ 時点前までの状態のみに依存する確率過程は $K$ 次マルコフ過程と呼ばれる．強化学習戦略は式 (3.2) が与える $K$ 次マルコフ過程として分析できる．

任意の $t$ に対し $X_t$ のとりうる状態系列の集合を $\mathcal{X}$ とする．履歴長 $K$ の強化学習戦略は，条件付確率 $P(X_{t+1}|X_t)$ を与え，初期状態によらず [3]，

[3] Perron-Frobenius の定理より，非周期的かつ既約な場合，定常分布が初期状態に依存しない．$K = \infty$ の場合，$\alpha_i = 1$ または $\beta_i = \infty$ において遷移行列が確率的ではなくなり，定常分布が初期状態に依存しうる．本論文では，近似的に有限の $K$ を用いたこと，および $\alpha_i = 1$ を分析対象から除外したことで，数値計算において周期的な解や既約可能な遷移行列は確認されなかった．

十分な時間経過ののち $|\mathcal{X}| = 4^K$ の状態系列上の定常分布へ収束する：

$$\pi(X) := \lim_{T\to\infty} P(X_0) \prod_{t=0}^{T} P(X_{t+1}|X_t) \ .$$

定常分布は遷移確率 $P(X_{t+1}|X_t)$ を用いて，$P(X_{t+1}) = P(X_t)$ の解として計算できる[4]．定常分布においては $P(X_{t+1}) = P(X_t)$ であり時点 $t$ に依存しないため，$K$ 次の履歴を $X := (x_1, x_2, \ldots, x_K)$ と表記する．

### 3.3.2 周辺確率

履歴長 $K$ の強化学習戦略に対し，$4^K$ の状態系列上の定常分布 $\pi(X) = \pi(x_1, x_2, \ldots, x_K)$ の和をとって周辺化し，1 状態の確率変数とすることで，$\mathcal{M} = \{\mathrm{CC}, \mathrm{CD}, \mathrm{DC}, \mathrm{DD}\}$ 上の周辺定常分布

$$\pi(\mathcal{M}) := \sum_{x_2,\ldots,x_K} \pi(x_1, x_2, \ldots, x_K)$$

を定義する．

本研究の目的すなわち個人と集団の意思決定における最適性を論じるためには，各戦略の相互協調の実現確率 $\pi(\mathrm{CC})$ および相互裏切の実現確率 $\pi(\mathrm{DD})$ を分析する必要がある．またすでに述べたように，1 次戦略は 1 回前の状態にのみ依存するため，特異的な戦略を除いて，周辺確率 $\pi(\mathcal{M})$ によって十分にその挙動が特徴づけられる．履歴長 $K > 1$ に基づく強化学習戦略のすべての性質の分析には周辺確率 $\pi(\mathcal{M})$ だけでは不十分であるが，1 次戦略との比較，および強化学習戦略による相互協調 $\pi(\mathrm{CC})$ および相互裏切 $\pi(\mathrm{DD})$ の定量的分析には十分である．このことから，本研究では周辺確率 $\pi(\mathcal{M})$ を分析した．

## 3.4 強化学習戦略による相互協調

IPD の枠組みにおいて，両プレイヤが個別に個人の利得を最大化した結果として，集団全体としても利得を最大化することが可能だろうか．この問いに対して，本節では主に強化学習戦略をとるプレイヤ同士の IPD

[4]定常分布 $P(X_{t+1}) = P(X_t)$ は遷移確率 $P(X_{t+1}|X_t)$ を表す行列の最大固有値に対する固有ベクトルとして求まる．

を分析する．前述のとおり，強化学習戦略は基本的に個人的な期待利得を最大化するよう過去の履歴に基づき行動を選択する．IPD の利得行列によると，両プレイヤの平均利得を最大化するのは相互協調 CC である．そこで，本節ではまず強化学習戦略同士の IPD において相互協調の発生確率が高くなる条件を特定し，個人と集団の意思決定の最適性が一致する場合があることを確認する．

とくに強化学習戦略において重要な役割を担うパラメータのひとつ，記憶保持率パラメータの役割について分析を行った．具体的には，プレイヤ 1 の記憶保持率を $\alpha_1$，プレイヤ 2 の記憶保持率を $\alpha_2$ とし，記憶保持率の大きさや異なる記憶保持率をもつプレイヤ間の定常状態における行動選択確率 $\pi(\mathcal{M})$ を分析した．記憶保持率 $\alpha_1 = \alpha_2 = 0$ やそれに十分に近い場合はほぼ等確率で協調 C または裏切 D を選択するランダム戦略と近くなる．その一方で $\alpha_1 = \alpha_2$ が共に大きい場合や $\alpha_1 \neq \alpha_2$ の場合において相互協調が発生するかを調べた．

### 3.4.1 記憶保持率の分析

強化学習戦略のパラメータを $\beta_i = 1.0$，$K = 10$ と設定し，各プレイヤの記憶保持率を $\alpha_i = 0.2,\ 0.4,\ 0.6,\ 0.8$（$i = 1,\ 2$）とした場合にえられる周辺定常分布を図 3.2 に示す．この図から，双方とも高い記憶保持率をもつとき（$\alpha_1 = 0.8$，$\alpha_2 = 0.8$），相互協調 CC が高い確率で実現されることがわかる．ところが，双方とも記憶保持率がそれほど高くないとき（$\alpha_1 = 0.6$，$\alpha_2 = 0.6$）には，相互裏切 DD のほうが相互協調 CC の確率より大きくなる．さらに，片方だけが高い記憶保持率をもつとき（$\alpha_1 = 0.8$，$\alpha_2 = 0.6$）では DD と DC が高い確率で実現される．またプレイヤ 1 と 2 の記憶保持率を交換した場合（$\alpha_1 = 0.6$，$\alpha_2 = 0.8$）では，DD と CD が高い確率で実現される．最後に，双方とも十分な記憶保持率をもたないとき（$\alpha_1 = 0.2$，$\alpha_2 = 0.2$）ではすべての行動組みがほぼ等確率で実現される．

以上の結果から，強化学習戦略同士の IPD では両プレイヤがともに十分な記憶保持率 $\alpha_i$ をもつことが相互協調を実現するうえで重要だと予想される．また，記憶保持率に差がある場合にはより高い記憶保持率をもつほうがより高い期待利得をえている．相互裏切から相互協調への変化は $\alpha_i = [0.6, 0.8]$ の範囲で生じている．

図 3.3 は $\alpha_i = 0.0, 0.01, \ldots, 0.98, 0.99$ の範囲で相互協調 CC と 相互裏

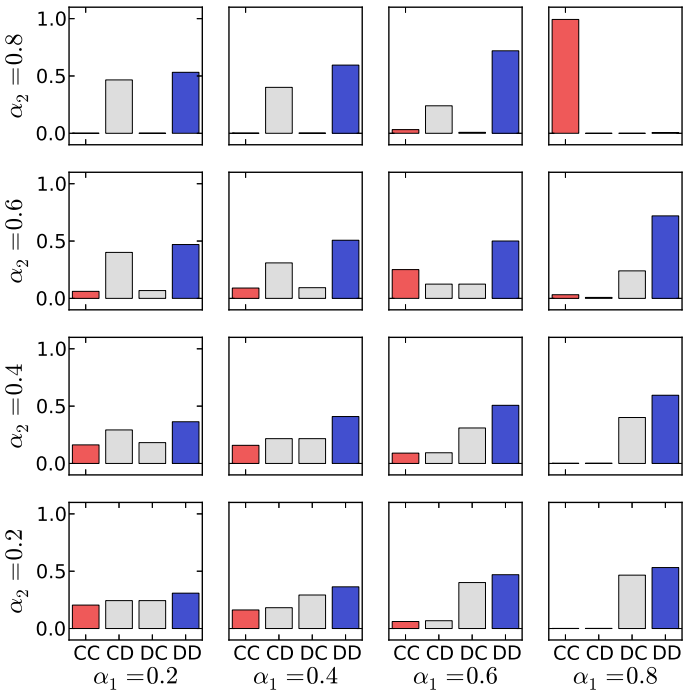



図 3.2: 記憶保持率 $\alpha_1$，$\alpha_2$ と周辺確率 $\pi(\mathrm{CC}), \pi(\mathrm{CD}), \pi(\mathrm{DC}), \pi(\mathrm{DD})$

切 DD の発生確率の差 $\pi(\mathrm{CC}) - \pi(\mathrm{DD})$ を示している．もし集団の最適性がより頻繁に実現されていれば $\pi(\mathrm{CC}) > \pi(\mathrm{DD})$ となる．この図から，DD に比べて CC が高い確率で実現されるには，厳密に等しい記憶保持率でなくてもよいが，双方が十分に高くかつ同程度の記憶保持率をもつことが重要であると考えられる．言い換えれば，この結果は一方だけが低い記憶保持率をもつ場合，相互裏切の確率がより高いことを意味する．

## 3.5 1 次戦略との比較分析

前節の分析では，強化学習戦略同士の囚人のジレンマにおいて相互協調 CC が発生することを確認した．またこの分析から，相互協調 CC の実現においては，学習のパラメータである記憶保持率が十分に高くかつ同程度であることが重要であると示された．他方で，これらの分析だけでは，強化学習戦略がどのように振る舞うことで相互協調を達成しているかは明らかではない．強化学習戦略は比較的長い行動履歴に基づき，また互いの行動に依存して学習し意思決定を行うため，その意思決定のダ

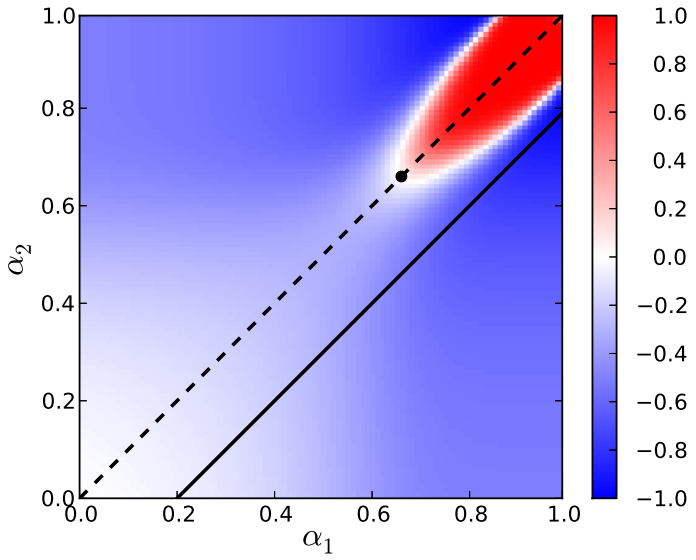



図 3.3: 相互協調確率と相互裏切確率の差

イナミクスは複雑である．そこで本節では IPD における各種の 1 次戦略の定常周辺分布と比較することで，強化学習戦略による相互協調の発生メカニズムを分析する．

### 3.5.1 定常周辺分布の分析

強化学習戦略と 1 次戦略を比較するために, 定常周辺分布 $(\pi(\mathrm{CC}), \pi(\mathrm{CD}), \pi(\mathrm{DC}), \pi(\mathrm{DD}))$ を分析する．定常周辺分布は $\pi(\mathrm{CC}) + \pi(\mathrm{CD}) + \pi(\mathrm{DC}) + \pi(\mathrm{DD}) = 1$ を満たすため，4 次元単体上に表現できる．また，今回の分析ではプレイヤの交換に対する対称性を考慮して $\pi(\mathrm{CD})$ と $\pi(\mathrm{DC})$ を区別しない．この場合，定常周辺分布は 3 次元単体にあり，平面 $(x, y) = (\pi(\mathrm{CC}) - \pi(\mathrm{DD}), \pi(\mathrm{CD}) + \pi(\mathrm{DC}))$ 上の点として一意に表現できる．図 3.4 では, 表 3.1 に示す 7 つの戦略のすべての組み合わせに対して定常周辺分布を算出し, この単体面上に白抜きの点 (○) で示した． 1 次戦略の複数の組み合わせが同じ点に重なるため, 白抜きの各点の側に 1 次戦略の組み合わせを表示している．図 3.4 に示す周辺分布の単体面は 3 つの頂点からなり，それぞれ (ALLC, ALLC)，(ALLD, ALLD)，(ALLC, ALLD) である．この図ではプレイヤの交換を同一視するため, (ALLC, ALLD) は (ALLD, ALLC) と同値である． また, 単体面の中央 $(0, 1/2)$ にランダム戦略 RAND

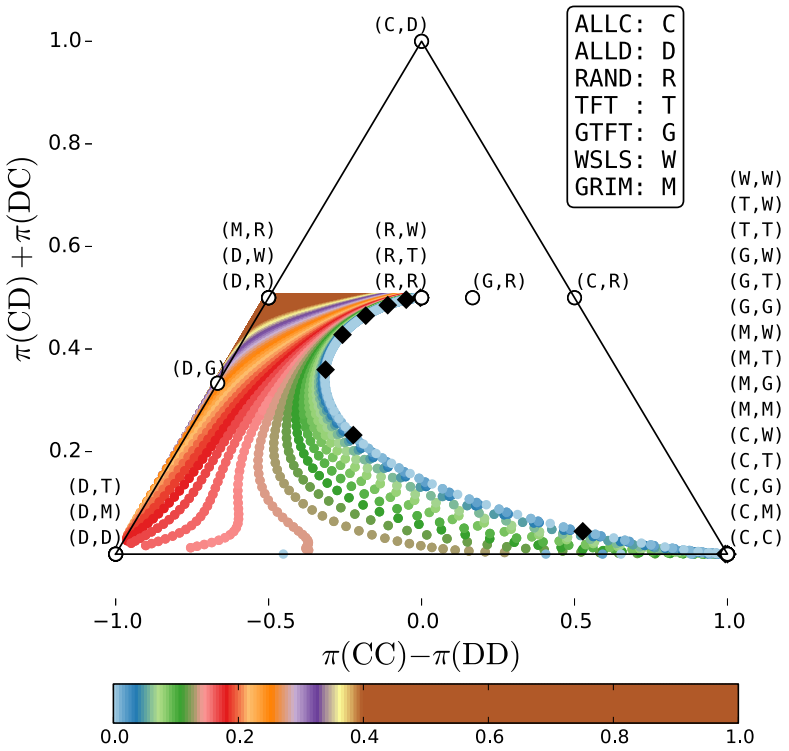



図 3.4: 定常周辺分布の単体面

がある．この単体面の 3 頂点，3 辺，または $\pi(\mathrm{CD}) + \pi(\mathrm{DC}) = 1/2$ 上に表 3.1 のすべての 1 次戦略ペアの結果が示されている．

### 3.5.2 強化学習戦略の 1 次戦略による近似

強化学習戦略の結果は縁無しの点（●）で示されている．単体面中央は (RAND, RAND) であるが，この点は強化学習における $\alpha_1 = \alpha_2 = 0$ の場合と一致する．強化学習戦略の結果は，$\alpha_i \to 1$ につれて単体面中央から離れ，単体面の辺上へと向かう．1 次戦略とは異なり，図 3.4 の単体面上の $\pi(\mathrm{CD}) + \pi(\mathrm{DC}) \leq 1/2$ に分布している．強化学習戦略の各点の色は両プレイヤの記憶保持率の差の絶対値 $|\alpha_1 - \alpha_2|$ から設定されており，青色の点では差がほぼ 0 で，緑色，赤色，橙色と変わるにつれて差が大きいことを表す．強化学習戦略が $\alpha_1 = \alpha_2 = 0.8$ あたりから (ALLC, ALLC) に漸近することを除いて，1 次戦略の分布から強化学習戦略の行動選択のメカニズムに関する洞察をえることは難しい．したがって，表 3.1 に

限らず，１次戦略クラスに含まれる戦略の中で，強化学習戦略の周辺確率分布をもっともよく近似する 1 次戦略を逆算し，この 1 次戦略の性質を分析した．

この分析では，同一の１次戦略同士の周辺定常確率が目標とする強化学習戦略の周辺定常確率と最小二乗誤差をもつよう遷移確率を定めた．１次戦略の定常分布 $(v(\mathrm{CC}), v(\mathrm{CD}), v(\mathrm{DC}), v(\mathrm{DD}))$ は１次マルコフ戦略 $(P(\mathrm{C}|\mathrm{CC}), P(\mathrm{C}|\mathrm{CD}), P(\mathrm{C}|\mathrm{DC}), P(\mathrm{C}|\mathrm{DD}))$ および初回に C をだす確率 $P_0(\mathrm{C})$ から定まる．そこで，これらの 5 変数を，強化学習戦略の１次周辺定常分布 $(\pi(\mathrm{CC}), \pi(\mathrm{CD}), \pi(\mathrm{DC}), \pi(\mathrm{DD}))$ を所与として，誤差関数 $\sum_{x\in\mathcal{M}}[\pi(x) - v(x)]^2$ を最小化するように求めた．強化学習戦略の周辺分布を近似する１次戦略は一般には一意に定まらず，無数にありえる．ここでは１次戦略のパラメータ空間上の一様乱数で 1000 個の初期値（5 変数）を与え，それらの初期値に対してえられた最小二乗解のうち，二乗誤差が十分に小さな閾値 $10^{-13}$ 以下の解の平均パラメータを分析した．

### 対称な強化学習戦略

図 3.4 の黒い菱形の点（◆）は，プレイヤ間で記憶保持率が等しい場合 $\alpha_1 = \alpha_2 = 0.1, 0.2, \ldots, 0.9$ の強化学習戦略同士の周辺分布（図 3.3 の破線）をもっともよく近似する１次戦略を示している．また，図 3.5 に近似１次戦略の遷移確率を $\alpha := \alpha_1 = \alpha_2$ の関数として示した．図 3.5 の各点は遷移確率の平均値，実線は $\alpha \pm 0.04$ の窓による移動平均を表す．１次戦略のパラメータは初期に協調 C をとる確率を含めて 5 つあるが，図 3.3 や図 3.4 から，$\alpha \geq 0.6$ の範囲では $\pi(\mathrm{CC})$ や $\pi(\mathrm{DD})$ が大部分を占め，他方，$\pi(\mathrm{CD})$ や $\pi(\mathrm{DC})$ が小さく，１次戦略のパラメータ $P(\mathrm{C}|\mathrm{CD}), P(\mathrm{C}|\mathrm{DC})$ の影響は弱いことが伺える．そこで，図 3.5 では $P(\mathrm{C}|\mathrm{CC}), P(\mathrm{C}|\mathrm{DD})$ のみを図示した．

図 3.5 から，もっともよく近似する１次戦略が $\alpha = 0.65$ 付近を境に切り替わっていることがわかる．記憶保持率 $\alpha < 0.65$ の範囲では $P(\mathrm{C}|\mathrm{CC})$ が低く，$P(\mathrm{C}|\mathrm{DD})$ が高い．一方で，$\alpha \geq 0.65$ の範囲では $P(\mathrm{C}|\mathrm{CC})$ が高く，$P(\mathrm{C}|\mathrm{DD})$ が低い．

図 3.5 の結果を，確率 $P(\mathrm{C}|\mathrm{CC})$ と $P(\mathrm{C}|\mathrm{DD})$ のパタンから，記憶保持率 $\alpha$ の区間に対応する 4 つのフェーズ I，II，III，IV に便宜的に分類して説明する．フェーズ I は $0.0 \leq \alpha < 0.3$ の区間に対応し，RAND（ランダム選択）と類似した戦略がみられる．フェーズ II は $0.3 \leq \alpha < 0.6$ の区間に対応し，前回の状態が CC でも DD でも裏切る確率のほうが高

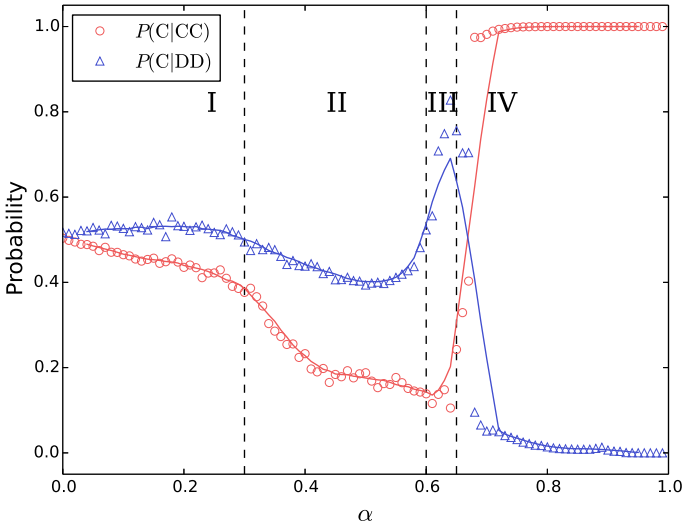



図 3.5: 対称な強化学習戦略（$\alpha_1 = \alpha_2$）を近似する 1 次戦略の C をとる条件付確率

いという点で確率的なゆらぎのある ALLD（いつでも裏切）に近い戦略であると解釈できる．

フェーズ III は $0.6 < \alpha < 0.65$ の区間に対応し，強化学習戦略は合理的プレイヤのように相互協調 CC から裏切に転じるが，一方で相互裏切 DD の場合には協調に転じる．先行研究で分析された代表的な戦略（表 3.1）にはこのような 1 次戦略は含まれていない．これらの代表的な戦略として解釈すれば，ALLD（相互協調時にも裏切る）と WSLS （相互裏切に対して協調する）を混合した戦略とみなせる．この戦略は，結果的に状況を相互協調を裏切へ，相互裏切を協調へと撹乱する戦略だと解釈できる．

最後に，フェーズ IV は $\alpha > 0.65$ の区間に対応し，強化学習戦略は TFT（しっぺ返し）戦略に類似した振る舞いを示すと解釈できる．換言すると，相互協調時にはそれを継続し，相手が裏切る場合には自分も裏切返す．実際にはいずれの場合でも強化学習戦略は直接的に相手の行動を参照して行動を選択していないが，累積利得を通じて間接的に TFT と類似の行動パタンを示すことが可能になったと考えられる．また $\alpha = 0.65$ 付近の前後で，こうした近似 1 次戦略の分岐が起こり，この分岐点は相互協調確率と相互裏切確率の差が 0 になる点（図 3.3 の破線上の黒丸の点 $\alpha_1 = \alpha_2 = 0.65$）と一致する．図 3.3 では $\alpha_1 = \alpha_2 = 0.65$ は裏切優

位の対戦結果から協調優位の対戦結果への分岐点であり，このことから図 3.5 ではこの分岐点の前後で裏切優位の戦略から協調優位の戦略へ変化したことを示している．$\alpha = 0.65$ の理論的な意味はまだわかっていないが，この数字は強化学習戦略が偶発的な行動を減らして相互協調で安定するために必要な累積利得の下限と対応していると思われる．

強化学習戦略では利得の順序関係だけでなく，その利得の大きさによって行動の選択確率が変化するため，IPD の満たすべき不等式の範囲で利得の値を変えて同様の分析を行った（3.6 節）．その結果，異なる利得行列の値においても，同様に強化学習戦略を近似する 1 次戦略の分岐点と相互協調の確率と相互裏切の確率が等しくなる点での $\alpha$ の一致が見られた．

以上の分析から，対称な強化学習戦略において相互裏切よりも相互協調が高い確率で発生した背景には，記憶保持率の低くゆらぎのある裏切戦略から，1 次戦略でいうしっぺ返し戦略（TFT）への振る舞いの定性的な変化があると考えられる．

### 非対称な強化学習戦略

次に 2 つの強化学習戦略が異なる記憶保持率 $\alpha_1 > \alpha_2$ をもつ場合を，対称な場合と同様に強化学習戦略を 1 次戦略で近似することにより分析した．すでに述べたとおり，記憶保持率に差がある場合，両プレイヤの記憶保持率が比較的高くても，相互協調より相互裏切の確率が高くなる．この分析では，同等の記憶保持率をもつ場合と比較し，強化学習戦略の振る舞いにどのような違いがあるかを調べる．具体的には，$0.0 \leq \alpha_2 < 0.8$ の範囲で $\alpha_1 = \alpha_2 + 0.2$ を満たす場合を，異なる記憶保持率をもつ強化学習戦略の典型的な場合として分析した．図 3.3 の実線で示されるとおり，この記憶保持率パラメータの場合，$\alpha_1$ が高いほど相互裏切の確率が高い．

図 3.6 は，記憶保持率の高い強化学習戦略（$\alpha_1$）と低い強化学習戦略（$\alpha_2$）を近似する 1 次戦略のパラメータをそれぞれ示す．図は両プレイヤの $P(\mathrm{C}|\mathrm{CC})$ および $P(\mathrm{C}|\mathrm{DD})$ を含み，移動平均線は破線（プレイヤ 1）と実線（プレイヤ 2）で表される．図から，$\alpha_1 = 0.5$ および $\alpha_2 = 0.5$ 付近において戦略の定性的な変化がみられるが，これは記憶保持率が小さくほぼランダムな戦略から，記憶保持率が十分に大きく過去の履歴に依存した選択を行う戦略への変化であると考えられる．上記の変化を除いて，どちらのプレイヤに関しても，記憶保持率 $\alpha_1$，$\alpha_2$ の関数として定性的な戦略の変化は見られず，分析の範囲でほぼ同様のパタンが見られた．これらの図では，図 3.5 にみられるような戦略の急激な変化はみら

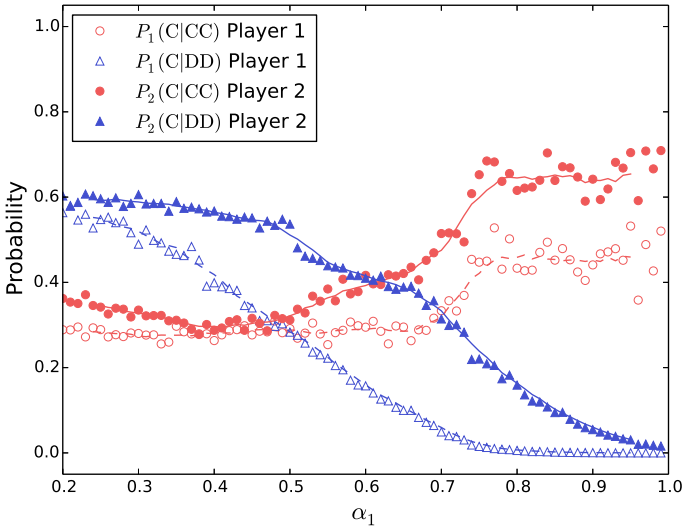



図 3.6: 非対称な強化学習戦略（$\alpha_1 = \alpha_2 + 0.2$）を近似する 1 次戦略の C をとる条件付確率

れない．また，いずれの記憶保持率の場合も $\alpha_i$（$i = 1, 2$）が大きいときには $P_i(\mathrm{C}|\mathrm{CC}) > P_i(\mathrm{C}|\mathrm{DD})$ の傾向がある．これは対称な強化学習戦略で $\alpha_1 = \alpha_2 > 0.65$ の場合（図 3.5）と定性的には類似している．一方，記憶保持率の異なる強化学習戦略のおもな違いは，相互に裏切った場合に次に協調する確率 $P(\mathrm{C}|\mathrm{DD})$ の違いにあると考えられる．$\alpha_1 > 0.5$ の範囲で，記憶保持率の低い強化学習戦略（図 3.6 実線）は，記憶保持率の高い戦略（図 3.6 破線）に比べて高い $P(\mathrm{C}|\mathrm{DD})$ を示しており，相互裏切 DD のあと協調 C へ転じやすい戦略となっている．非対称な強化学習戦略では相互裏切の周辺確率 $\pi(\mathrm{DD})$ が高いため，高い頻度で記憶保持率の低いプレイヤが協調し（相対的に高い $P_2(\mathrm{C}|\mathrm{DD})$），一方，記憶保持率の高いプレイヤが裏切る（相対的に低い $P_1(\mathrm{C}|\mathrm{DD})$）．したがって，高頻度で $\pi(\mathrm{DC})$ が発生し，$\pi(\mathrm{DD})$ が安定的に高い確率をもつと考えられる．

同程度の記憶保持率をもつ（対称な）強化学習戦略同士の場合に比べて，異なる記憶保持率をもつ（非対称な）強化学習戦略の間では「撹乱」的な行動から生じる協調行動が同期しにくく，相互裏切から相互協調へと転じる確率が低くなる．その結果，記憶保持率が同程度の場合とは異なり，両プレイヤが類似の行動選択をするにもかかわらず相互協調は不安定であり，相互に裏切るか，記憶保持率の高いプレイヤによる一方的裏切という状態のみが高い確率で発生すると考えられる．

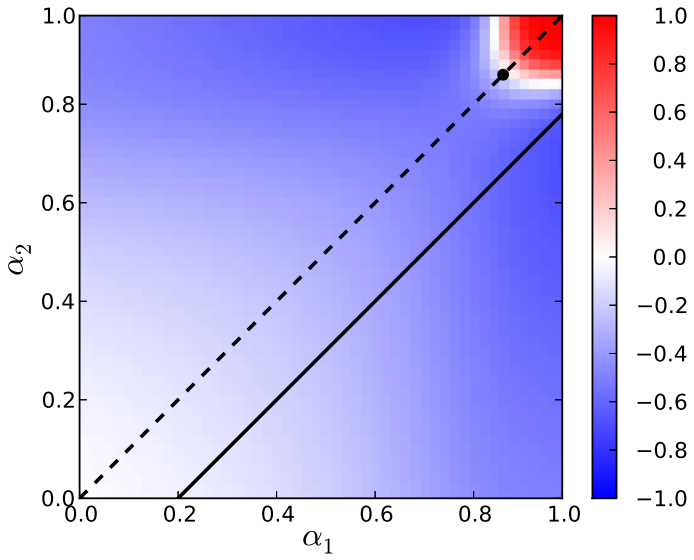



図 3.7: 図 3.3 に対応した異なる利得行列の結果

## 3.6 異なる利得行列を用いた分析

囚人のジレンマは利得の順序関係だけで定義されているが，強化学習戦略は利得の数値に基づいて意思決定を行うため，その利得の大きさによって行動の選択確率が変化する．そのため，本論文の結論の一般性を担保するには，異なる利得行列を用いた分析においても定性的に同じ結果がえられるかを調べる必要がある．本節では負の値を含む利得行列として，$(f(\mathrm{CC}), f(\mathrm{CD}), f(\mathrm{DC}), f(\mathrm{DD})) = (1,\ -2,\ 2,\ 0)$ として同様の分析を行った．強化学習戦略のパラメータは $\alpha_i = 0.01, 0.03, \ldots, 0.97, 0.99$, $\beta_i = 1.0$，$K = 10$ とした．本節ではふたつの利得行列を明示的に区別して記すために，利得行列 $(3,\ 0,\ 5,\ 1)$ を 3051，また $(1,\ -2,\ 2,\ 0)$ を 1220 を表記する．

図 3.7 は図 3.3 に対応する結果である．ふたつの図は定性的には類似した傾向を示しているが，利得行列 1220 では 3051 に比べてより高い $\alpha$ でなければ $\pi(\mathrm{CC}) > \pi(\mathrm{DD})$ とならない．その要因のひとつは利得行列の数値が全体的に小さく，さらに負の値を含むことにあると考えられる．また，対称なときには CC 優位であるが非対称な記憶保持率（$\alpha_1 \neq \alpha_2$）のときには DD 優位となるような点はみられない．これは利得行列 1220 がもつ利得の関係 $f(\mathrm{CD}) + f(\mathrm{DC}) = 0$ と関連すると推測される．

図 3.8 は図 3.5 に対応する結果である．利得行列 3051 の結果と同様

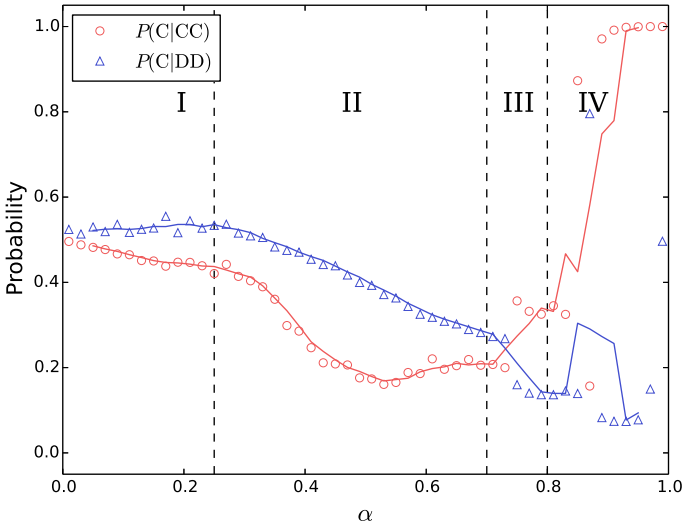



図 3.8: 図 3.5 に対応した異なる利得行列の結果

に，大別して 4 つのフェーズに分類できる．フェーズ I（$0.0 < \alpha < 0.25$）は RAND に近い戦略，フェーズ II（$0.25 < \alpha < 0.7$）は RAND と ALLD の混合された戦略，さらにフェーズ IV（$0.8 < \alpha < 1.0$）は TFT，GTFT に近い戦略として解釈でき，これらはいずれも 3051 の結果と整合的である．これに対して，フェーズ III（$0.7 < \alpha < 0.8$）は 3051 の結果と異なり，$P(\mathrm{C}|\mathrm{CC}) > P(\mathrm{C}|\mathrm{DD})$ となっている．これは確率的ゆらぎのより大きいランダムの高いフェーズ IV と解釈できるだろう．$\alpha = 0.85$ 付近を分岐点として，戦略の大きな変化がみられ，このときの $\alpha$ の値は図 3.7 において $\pi(\mathrm{CC}) > \pi(\mathrm{DD})$ へと切り替わる点と一致している．この一致は利得行列 3051 でもみられた．

図は示していないが，記憶保持率が非対称 $\alpha_1 > \alpha_2$ の場合でも，利得行列 3051 の結果と同様に，低い記憶保持率（$\alpha_2$）をもつプレイヤ 2 のほうが DD の状態において協調 C をだす確率が高い．したがって，利得行列 1220 においても，記憶保持率の高いプレイヤが記憶保持率の低いプレイヤよりも平均的に高い利得をえている．

## 3.7　強化学習戦略の１次戦略に対する振る舞い

本節では，強化学習戦略が個人の期待利得を最大化する戦略であることを確認する．そのため，表 3.1 の各１次戦略に対してほぼ同等かあるいは高い期待利得をえられるかを分析した．この分析では，強化学習プレイヤのパラメータは $\alpha = 0.8$，$\beta = 1.0$，$K = 9$ に設定した．これは今回のパラメータ範囲のなかでは，１次戦略との対戦において期待利得の最大化という目標を十分に達成できるパラメータ値である．定常分布を数値計算する際，初期値は各戦略の性質から成立しうる状態の集合上の一様分布とした．たとえば，強化学習と ALLC の対戦では ALLC が C しかとらないため，ALLC 側に D を含む状態の初期確率はすべて 0 とし，他方，RAND との対戦ではすべての状態の集合上の一様分布を初期値とした．以下では強化学習戦略を RL と略記する．

- **RL 対 ALLC**

  RL とのゲームの帰結は DC となる．これは RL にとって $f(\mathrm{DC}) > f(\mathrm{CC})$ から導かれる．RL の習得した戦略は ALLD とよく似ている．期待利得は RL のほうが高い．

- **RL 対 ALLD**

  RL とのゲームの帰結は DD となる．これは RL にとって $f(\mathrm{DD}) > f(\mathrm{CD})$ から導かれる．RL の習得した戦略は ALLD とよく似ている．期待利得は ALLD のほうがわずかに高い．これは RL が非常に小さい確率で C をだしうるためである．

- **RL 対 RAND**

  RL とのゲームの帰結は DD と DC を等確率でとる．これは RL にとって $f(\mathrm{CC}) + f(\mathrm{CD}) < f(\mathrm{DD}) + f(\mathrm{DC})$ から導かれる．RL の習得した戦略は ALLD とよく似ている．期待利得は RL のほうが高い．

- **RL 対 GRIM**

  RL とのゲームの帰結は DD となる．これは RL が確率的に振る舞うため，他の１次戦略のように決定的に C をだせないという事実から導かれる．RL の習得した戦略は ALLD とよく似ている．期待利得はほぼ同じである．

- **RL 対 TFT**

  RL とのゲームの帰結は CC となる．これは RL にとって $f(\mathrm{CC}) + f(\mathrm{CC}) > f(\mathrm{DC}) + f(\mathrm{CD})$ から導かれる．RL の習得した戦略は ALLC とよく似ている．期待利得はほぼ同じである．

- **RL 対 GTFT**

  TFT と同じ理由で繰り返しゲームの帰結は CC となる．しかし，TFT に比べて，DD が減り，代わりに DC が増えている．これは GTFT の確率的に C をだすという方針から導かれる．RL の習得した戦略は ALLC とよく似ている．期待利得はほぼ同じである．

- **RL 対 WSLS**

  RL とのゲームの帰結は DD と DC を等確率でとり，DD → DC → DD → を繰り返す．これは RL にとって $f(\mathrm{DD}) + f(\mathrm{DC}) > f(\mathrm{CC}) + f(\mathrm{CD})$ であることに加えて，WSLS の反対の行動を選ぶという方針から導かれる．RL の習得した戦略は ALLD とよく似ている．期待利得は RL のほうが高い．

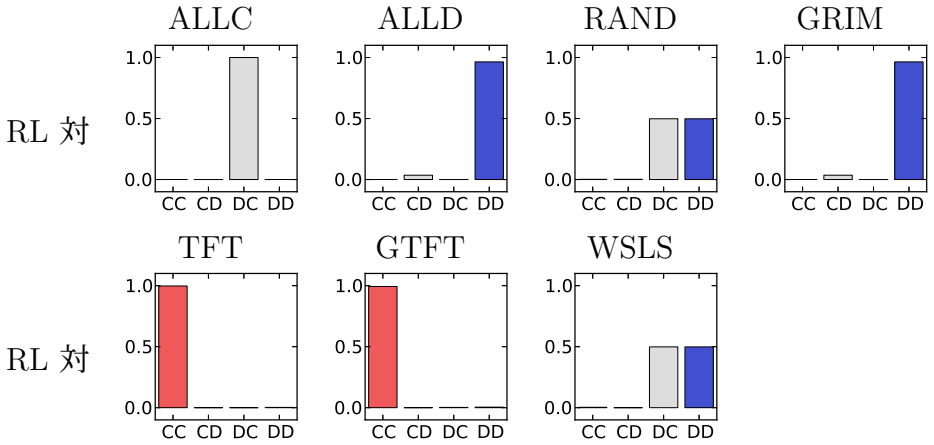
### 3.7.1 まとめ

以上の結果を表 3.2 に要約する．RL はいずれの 1 次戦略との対戦においても，自分の期待利得を最大化し，その結果，対戦相手より高い期待利得をえるか，ほぼ同じ期待利得をえている．強化学習が ALLC のように振る舞うのは TFT，GTFT に対してのみで，これらの戦略との対戦においてのみ強化学習は CC を実現する．その他の戦略には，強化学習は ALLD のように振る舞う．ALLC，RAND，WSLS など DD で安定しない戦略は RL より低い期待利得しかえられない．RL はあくまでも確率的に振る舞うため，ALLD のような常に D をだす戦略に対してはほぼ同等の期待利得をえるものの，わずかに劣る．

このように，強化学習戦略は同じ戦略を採用する相手プレイヤと相互協調できるだけでなく，進化ゲームの文脈で登場する 1 次戦略に対してもほぼ最適に振る舞うことができている．

表 3.2: 強化学習（RL）対 1 次戦略 の帰結

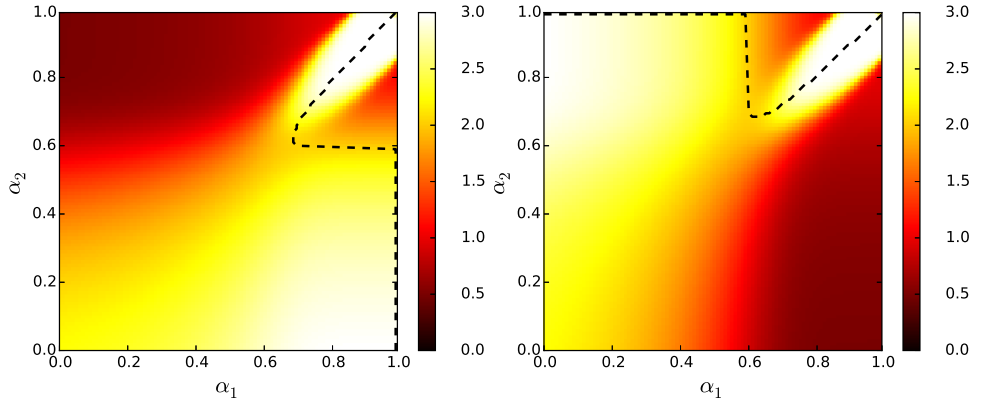
| | 相手 | | 学習結果 |
|---|---|---|---|
| RL 対 | ALLC | いつでも C | ALLD |
| | ALLD | いつでも D | ALLD |
| | RAND | C と D を等確率 | ALLD |
| | GRIM | 一度でも相手が D をだすまでは C | ALLD |
| | TFT | 相手の直前の行動を模倣（最初 C） | ALLC |
| | GTFT | TFT 亜種．相手が D でも確率 1/3 で C | ALLC |
| | WSLS | 負けなら行動を切り替え，勝ちなら同じ | ALLD |




## 3.8 強化学習戦略の“進化”

本章ではこれまで，両プレイヤがある強化学習戦略 $\alpha_i \in [0,1]$ を選択したときの IPD の帰結を分析してきた．本節では分析の焦点を戦略からプレイヤに移し，両プレイヤが IPD ゲームの結果を踏まえて，自己利益を最大化するという目的のもと強化学習戦略を切り換えるゲーム [5] を考える．これはゲーム理論の用語では Nash 均衡となる戦略ペアを求めることに対応する．このゲームでは各プレイヤは $\alpha_i \in [0,1]$ を変化させることで同時に（交互に）最適な強化学習戦略を求めるが，これは相手の戦略に対して最適な記憶保持率を求めることを意味する．したがって，本節では，両プレイヤが自己利益を最大化するよう戦略を変化させた場合，最終的にどのような強化学習戦略（記憶保持率）を採用するかを調べる．

[5] このゲームは進化ゲームと類似しているが，進化ゲームのプレイヤは「自然」1 人であるのに対し，本節では 2 人のプレイヤを想定している．




Player 1 $F_1(\alpha_1, \alpha_2)$　　Player 2 $F_2(\alpha_1, \alpha_2)$

図 3.9: 強化学習戦略 $\alpha_i \in [0, 1]$ を扱うプレイヤの利得関数 $F_i(\alpha_1, \alpha_2)$

図 3.9 では図 3.3 と同じデータを用いて各戦略ペアごとの期待利得を示した．より具体的には，図 3.3 の各点 $(\alpha_1, \alpha_2)$ に対応した定常分布 $(\pi(\mathrm{CC}), \pi(\mathrm{CD}), \pi(\mathrm{DC}), \pi(\mathrm{DD}))$ から，$f_1(\mathrm{CC})\pi(\mathrm{CC}) + f_1(\mathrm{CD})\pi(\mathrm{CD}) + f_1(\mathrm{DC})\pi(\mathrm{DC}) + f_1(\mathrm{DD})\pi(\mathrm{DD})$ を計算し，これを図 3.9 の各点 $(\alpha_1, \alpha_2)$ の値とした．これは強化学習戦略 $\alpha_i \in [0, 1]$ を戦略集合とするゲームの利得関数 $F_i(\alpha_1, \alpha_2)$ を近似するものといえる．強化学習戦略 $\alpha_i = 0$ はランダムに振る舞うため $(\alpha_1, \alpha_2) = (1, 0)$ は ALLD vs RAND のようになり，プレイヤ 1 の期待利得は $F_1(1, 0) = [f_1(\mathrm{DC}) + f_1(\mathrm{DD})]/2 = (5+1)/2 = 3$ となり，他方，プレイヤ 2 は $F_2(1, 0) = (0 + 1)/2 = 0.5$ となる．これは相互協調の期待利得 $f_1(\mathrm{CC}) = 3$ と等しい．また，$(\alpha_1, \alpha_2) = (0, 0)$ では期待利得は $(3 + 0 + 5 + 1)/4 = 2.25$ である．

Nash 均衡となる戦略ペアは互いに相手の戦略に対する最適応答となるが，これは両プレイヤの利得関数を同時に最大化する極大点と対応する．図 3.9 中，破線が相手の戦略に対する最適応答を表す．図から，強化学習戦略を集合とするゲームの利得関数 $F_1$（$F_2$）は 2 つの極大点をもつと推測される．一方は $(\alpha_1, \alpha_2) = (1, 0)$ 付近であり，この点はプレイヤ 1 の利得関数 $F_1$ では極大であるが，プレイヤ 2 の利得関数 $F_2$ では極大ではない．他方は相互協調 $(\alpha_1, \alpha_2) = (1, 1)$ 付近であり，この点は両プレイヤにとって極大となる．最適応答を示す破線は $\alpha_1 = \alpha_2 > 0.65$ の範囲では $F_1$ と $F_2$ でほぼ一致しており，$\alpha_i \to 1$ で完全に一致すると予想される．他方，$\alpha_i < 0.65$ では一致していないものの，大部分において最適応答は $\alpha_i = 1$ を示している．以上から，強化学習戦略を最適化するプレイヤは

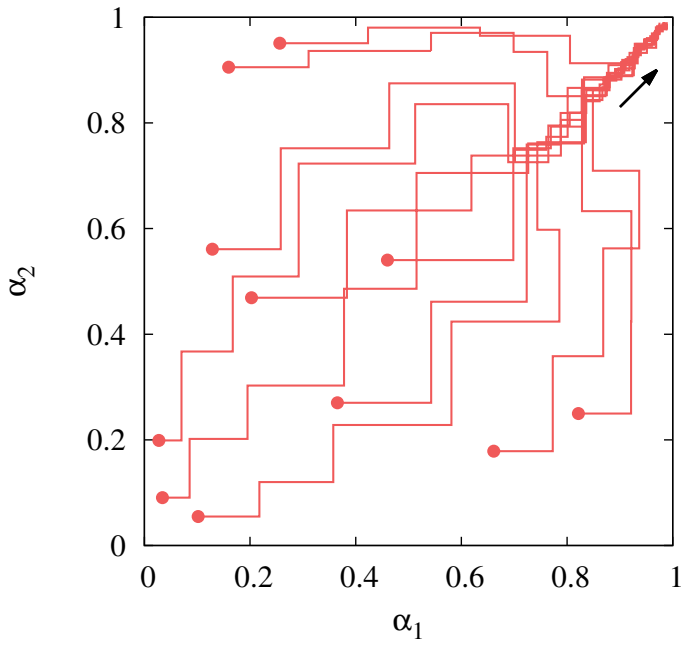



図 3.10: プレイヤが $\alpha_i$ を探索し利得関数 $F_i$ を交互に最大化する様子（各 ● は異なる初期値 $(\alpha_1, \alpha_2)$）

最終的に $\alpha_i = 1$ を選択することが予想される．また，$(\alpha_1, \alpha_2) = (1, 1)$ では強化学習戦略のゲームの唯一の Nash 均衡であると考えられる．この唯一の均衡では，図 3.3 より，強化学習戦略は相互協調 CC をほぼ確率 1 で実現できる．

以上の予想をシミュレーションにより検証する．図 3.10 は，複数の異なる初期値（図中 ●）から，各プレイヤが交互に期待利得 $F_i$ を最適化する様子を示す[6]．図 3.10 では最適化の経緯を捉えるため，各プレイヤには戦略を局所的に探索させた（$\alpha_i \in [0, 1]$ 全域を探索させても最適化の結果は同じ）．図から，交互に局所的に最適化した場合，初期値によらず，記憶保持率 $\alpha_i \to 1$ すなわちより長期的な履歴を考慮する強化学習戦略が次第に選ばれることがわかる．換言すれば，図 3.9 から推測される，強化学習戦略を戦略集合とするゲームの唯一の Nash 均衡へ収束している．

[6]プレイヤ $i$ は現在の戦略 $\alpha_i$ の近傍 $\{\alpha_i' : \alpha_i' \in [\alpha_i - \varepsilon, \alpha_i + \varepsilon] \subset [0, 1]\}$ をモンテカルロ探索し，そのうち相手の戦略 $\alpha_j$ に対して最大の期待利得 $F_i(\alpha_i^*, \alpha_j) \geq F_i(\alpha_i', \alpha_j)$ を与える戦略 $\alpha_i^*$ を新しい戦略として採用する．これをプレイヤ 1，2 交互に繰り返す．

## 3.9 議論

### 3.9.1 まとめ

本研究では，個人と集団の利害の対立を表現するモデルのひとつである繰り返し囚人のジレンマを題材に，不確実な状況下での意思決定による利害解消の可能性について論じてきた．古典的な分析では，囚人のジレンマにおける合理的な帰結（Nash 均衡）は相互裏切であるとされ，それを避けて相互協調（Pareto 効率的な解）を実現可能にする条件について議論されてきた．本研究では，互いに相手の行動に関する情報を与えられず，しかし個人の利得を最大化する強化学習戦略を考え，実効的に同程度の長さの行動履歴に基づき行動を選択する強化学習戦略の間では，相互協調が高い確率で発生することを示した．また，1 次戦略による近似から，この相互協調の発生は学習によるしっぺ返し戦略のような戦略の発生に伴って起こることも明らかになった．

### 3.9.2 他のアプローチとの関係

これまでに，相互協調の実現は Axelrod [5] や Nowak and Sigmund [45, 46] に代表される進化的手法により詳しく分析されてきた．これらの研究で対象とされた代表的な 1 次戦略（TFT など表 3.1 参照）は直前の相手の行動に依存した戦略であり，それぞれの戦略（個体）の目標は与えられた集団のなかでより高い平均利得（適応度）をえることに置き換えられている．つまり，これらの先行研究は，個人（個体）の期待利得の最大化ではなく，集団内の他の戦略に対して適応的に行動を選択する問題として IPD を解釈して，相互協調の実現可能性を論じている．したがって，進化的な問題として論じられてきた IPD における相互協調は，古典ゲーム理論および本研究で対象とするジレンマとは異なる問題である．

古典ゲーム理論 [63] における合理的主体は，自分と相手の行動組みの関数である利得行列を参照して自己利益を最大化する行動を選択する．繰り返しゲームにおいても合理的主体は同じ論理で意思決定を行う．合理的主体同士のゲームは個人合理的な解（相互裏切）へ帰結される．古典的な枠組みにおいて学習を導入したモデルとして信念学習 [12] がある．信念学習は相手の行動に対する利得に基づいて行動の学習を行う．合理的主体と信念学習主体は後方帰納か逐次学習かという点で異なるが，この違いにもかかわらず，信念学習を行うプレイヤ同士のゲームは個人合理

表 3.3: 囚人のジレンマへのアプローチ

| | 相手の選択 | 利得行列 | 最大化 |
|---|---|---|---|
| 古典 | — | 既知 | 自己利益 |
| 信念学習 | 既知 | 既知 | 自己利益 |
| 進化 | 既知 | 既知 | 適応度 |
| 強化学習 | 未知 | 未知 | 自己利益 |

的な解（相互裏切）へ帰着し，集団合理的な解（相互協調）は実現されない．

### 3.9.3 不確実性のある意思決定

以上の先行研究の枠組みおよびその結論の概略を表 3.3 に要約した．これらの先行研究では，十分な情報に基づく強い推論によって，個人の合理性を超えた相互協調は発生せず（古典ゲーム理論，信念学習），多様な戦略への適応という進化的な視点を導入することで相互協調の発生を模索してきた（進化的分析）．一方，本研究では基礎的な囚人のジレンマの利得構造に対し，個々のプレイヤが自己の期待利得を最大化することがそのまま集団としての期待利得を最大にする場合がありえるか，という観点から分析を行った．

本研究で分析した強化学習戦略は自分の行動のみに対する利得に基づき意思決定し，自己利益を最大化する．強化学習主体にとって，相手の行動や，自分と相手の行動の関数としての利得行列は不可視であり，自分の行動の関数として不確実な利得のみが与えられる．利得が確率的で不確実であることが，強化学習同士の対戦において，個人と集団のジレンマの解消，つまり個人と集団の利得の最大化が一致した主要因であると考えられる．

本研究で IPD に新しく導入した要因は，利得が相手の行動によるという知識の欠如である．一般に囚人のジレンマのように行動と帰結のすべてが事前に明らかであるような状況は少なく，むしろ特定の問題に関する利害構造はその問題に取り組んだ結果から推測しなければならないことが多いと考えられる．この観点からは，ゲーム理論的な合理性よりも，むしろ自分の過去にとった行動とそれに対する利得のみから，統計的により期待利得の高い行動を選ぶ強化学習戦略は自然であると考えられる．

本分析では，このような意思決定に関する情報の不完備性（それに伴う不確実性）を導入した結果，繰り返し囚人のジレンマのように個人と集団の合理性が対立するゲームにおいて，個別に利益を追求することと集団での利益の最大化が一致する場合があることを示した．

### 3.9.4 強化学習を用いた IPD 研究との関連性

本論文と同じく，強化学習を IPD に適用した研究 [6, 37, 50] では，相互協調は実現されていないか [50]，自分の行動を相手の行動と関連づけて学習した場合のみ相互協調がある程度高い確率で観察される [6, 37] ことが示されている．先行研究の結果は「相手の行動が見える」という点で本論文よりも不確実性の少ない意思決定といえる．先行研究において自分に関する情報だけの場合に相互協調が観察されなかったのは，定常分布を分析することの技術的な難しさにあると思われる．先行研究はエージェントシミュレーションを用いているが，その場合，記憶保持率 $\alpha_i$ が 1 に近づき，より長期的な履歴から学習するほど，その定常分布を調べるには一般にシミュレーションをより長期間走らせ，より数多くのゲームをプレイさせる必要がある．ところが，これらの先行研究は十分なゲーム回数を実行していない可能性がある．著者が $\varepsilon$-greedy TD 学習のエージェントシミュレーション [37] を追試したところ，約 100 倍ほどのゲーム回数を行えば，本論文の主張と同じように，自分に関する情報だけの場合でも相互協調が実現できることがわかっている．

### 3.9.5 学習の結果としてのしっぺ返し行動の発現

3.8 節では，強化学習の結果としてしっぺ返し行動が発現することを論じたが，他方，進化ゲーム理論でも自然選択の結果としてしっぺ返し戦略が発現することが知られている．強化学習戦略からしっぺ返し戦略が発現することは「低次の戦略として観察した場合にそうみえる」ということであり，それはプレイヤがある強化学習戦略を選択した結果である．他方，進化ゲームでは１次戦略そのものを自然選択する．強化学習と自然選択は何らかの最適化を実行しているという点で共通しているが，しっぺ返し戦略の発現メカニズムは根本的に異なる [7]．

[7] Borgers and Sarin [8, 9] は学習と進化の本質的類似性を論じているが，不自然な仮定のもとで学習と進化の類似性を示しており議論の余地がある．

### 3.9.6 実験的な知見との関連性

本研究の分析結果は，繰り返し囚人のジレンマの利得構造を知らないプレイヤ同士が，互いの選択に依存する利得に基づいて学習を行う場合に至る行動選択の確率を示している．本研究の想定する状況は，(a) 無限に繰り返すゲームにおいて，(b) 利得が未知で同一の選択に対して複数の利得が確率的に割り当てられる状況である．本研究の想定するこうした状況に関連する心理学・行動経済学の知見からは，(a) 繰り返しゲームの回数に関する不確実性 [7] や (b) ゲームの利得に関する不確実性 [28, 29] が人の協調行動の選択確率に影響することが示されている．具体的には，期待されるゲームの繰り返し回数が多いほど，ゲームの開始時に協調する確率が高くなることが示されている [7]．また，協調が必ずしも裏切より不利ではないサブゲームを含み，それらをランダムにプレイするような不確実なゲームにおいても，協調の確率が高まることが報告されている [28, 29]．これらの知見は本研究の想定する状況と部分的に類似したゲームにおいて人が協調する可能性を示しており，本研究の分析結果とも一貫性がある．

### 3.9.7 今後の課題

これらの実験的な知見は，ごく短期的なゲームにおける意思決定を反映しており，これだけでは実社会で見られる人の長期的かつ安定的な協調行動を説明することは難しい．この点で，本研究の理論的な分析には特定の戦略による長期的な行動選択の確率（定常分布）を調べられるという利点がある．一方，本研究からえられる知見は理論的な予想に留まっており，現実の人の意思決定への示唆としては限界がある．最後に，実験・理論の両面から人の相互協調のメカニズムに迫るために，本研究の理論的な示唆をいかに実験的に検証できるかについて今後の見通しを示したい．

本分析で着目した記憶保持率は，行動の選択確率がどの程度長期的に過去のゲームの結果に依存するかを表す．図 2 で示したとおり，両プレイヤの記憶保持率の組み合わせにより相互協調の確率が決まり，また図 4 では記憶保持率の違いによって異なる 1 次戦略で近似できることを示した．

こうした理論上の記憶保持率と，実験における人の戦略を対応づけるには，主に 2 つの方法が考えられる．1 つ目は，被験者の行動選択のデータに基づき，それを最も良く説明する理論的な戦略を推定することであ

る．2 つ目は，記憶保持率と想定される認知処理を促進または妨害し，その要因と相関するゲームにおける戦略の変化を調べることである．

この 2 つの方法を併せた研究 [38] では，二重課題によって作業記憶を制限した場合，特定の 1 次戦略として分類した人の戦略が変わりうることが示唆されている．この課題は本研究の想定するすべての条件を満たしてはいないが，モデル上で定義される記憶保持率の解釈に示唆的である．つまり，Milinski and Wedekind [38] の研究を踏まえると，提案モデルの記憶保持率は，未知なるゲームにおいて意思決定をするプレイヤの作業記憶に相当すると予想できる．想定される実験において，各プレイヤは各行動をとったときの平均的な利得の情報を保持・更新する必要があり，これに関わる作業記憶の妨害によって，対応するモデル上の記憶保持率を低下させる効果をもつと期待される．近年では，囚人のジレンマなどのゲームを行う行動主体の学習モデルを推定する研究も行われている [60, 61, 65]．今後の課題として，この理論的な予想を実験的に検証する研究が考えられる．

# 4　強化学習戦略を扱うゲームのベクトル場近似

## 4.1　本章の目的

学習主体のゲームを解析的に分析しようという試みがある．エージェントシミュレーションや前章の有限マルコフ過程を用いた分析は確率過程の定常分布を知ることができるが，反面，存在する解の集合，解の存在条件，解の安定性などを解析的に調べることが困難である．ゲーム理論の研究では，学習主体のゲームを微分方程式として近似し，その性質が調べられている．本章では強化学習戦略の囚人のジレンマを確率差分方程式として近似し，その平均的挙動のベクトル場を調べることで，存在する解の集合，解の存在条件，解の安定性などを解析する．

## 4.2　学習主体のゲームの近似

Bush and Mosteller [13, 14] は要求学習を確率モデルとして定式化した．その平均的挙動は理論的に解析され [20]，この結果を踏まえた Borgers and Sarin [8] は要求レベルの項を無視した場合（含めた場合は [9]），連続時間の極限（時定数 $\to 0$）においてこの確率モデルの平均的挙動は連続時間結合レプリケータと一致することを指摘した．Borgers and Sarin [8] は要求学習の力学的近似モデルと見なされている．他方，Sato and Crutchfield [51] や Tuyls et al. [59] は強化学習のひとつである Q 学習を用い，[8] と同じように連続時間結合レプリケータを導出した．この連続時間結合レプリケータを用いた囚人のジレンマの研究 [31] は，相互裏切が唯一の解であることを理論的に示した．また，期待利得を勾配とした勾配降下法による相互学習の力学的アルゴリズム [55]（IGA アルゴリズム）は相互裏切が唯一の解であることを数値計算により示した．また，IGA アルゴリズムの拡張型として WoLF 原理 [10] や GIGA アルゴリズム [66] など

があり，アルゴリズムの理論的性質や Nash 均衡解への収束安定性などが議論されている．これらは信念学習の力学的近似モデルとみなせる．

いずれの研究も相互学習のダイナミクスおよびそのアトラクタを調べることを目的とするが，本来の確率モデルを力学モデルで近似するさいにいくつかの仮定を置いており，極限でのダイナミクスが本来のモデルとは一致しない可能性がある．また，ときに不自然な仮定を置いているにもかかわらず，近似モデルがどういう点で本来のモデルを近似できているかは検証されないことが多い．実際のところ，連続時間結合レプリケータ [51] のアトラクタ上での平均的振る舞いはエージェントシミュレーションの平均的振る舞いと一致しないことが報告されている [26].

本論文は相互学習による協調問題を扱うが，個別のダイナミクスよりむしろ，定常状態あるいは協調問題の解決可能性を問題としてきた．この問題意識のため，本論文ではダイナミクスを扱う先行研究とは異なる仕方で近似モデルを定式化する．本章では，この近似モデルを用いて，存在する解の集合，解の存在条件，解の安定性などを解析する．さらに，近似モデル（本章）の理論解析の結果が本来のモデル（前章）の結果とどのような点で定性的に類似するかを確認する．

## 4.3 定式化

### 4.3.1 強化学習のベクトル場近似

強化学習戦略では，行動 $x$ に対する割引された累積利得

$$R_{i,x}(X_t) = \sum_{k=1} \alpha_i^k \, \delta(x, x_{i,t-k}) \, f_i(x_{t-k}) \tag{4.1}$$

に基づき，プレイヤ $i$ は次式の確率で行動 $x$ を選択する：

$$P_i(x|X_t) = \frac{\exp[\beta_i \, R_{i,x}(X_t)]}{\sum_y \exp[\beta_i \, R_{i,y}(X_t)]}$$

ここで, $X_t = (x_{t-1}, \ldots, x_1)$ は時点 $t$ までの行動ペアの履歴, $\alpha_i$ は割引率パラメータ, $\beta_i$ は鋭敏性パラメータである．囚人のジレンマでは $x \in \{\mathrm{C}, \mathrm{D}\}$ だから，累積利得の差 $\Delta R_i(X_t) = R_{i,\mathrm{C}}(X_t) - R_{i,\mathrm{D}}(X_t)$ を用いて，

$$P_i(\mathrm{C}|X_t) = \frac{1}{1 + \exp[-\beta_i \Delta R_i(X_t)]}$$
$$P_i(\mathrm{D}|X_t) = 1 - P_i(\mathrm{C}|X_t)$$

と書ける．すなわち，選択肢の数 $M = 2$ のゲームではパラメータと累積利得の差が与えられれば，行動確率は一意に定まる．ここで，累積利得の差 $\Delta R_i(X_t)$ は履歴 $X_t$ に依存して離散的に変化するが，本節の近似モデルではすべての $\Delta R_i(X_t) \in [-\infty, +\infty]$ を分析対象とする代わりに，履歴そのものを考えない．これは個別のダイナミクスを大域的に追いかけるアプローチに対して，ある時点での局所的な振る舞いのみを分析対象とすることで，近似的に均衡解を調べるアプローチを意味する．以降では，履歴 $X_t$ を省略し，$\Delta R_i$，$P_i(\mathrm{C})$，$P_i(\mathrm{D})$ と記す．また，$P_i(\mathrm{C}) = 1 - P_i(\mathrm{D})$ であるから，$P_i := P_i(\mathrm{C})$ と記す．この表記を用いれば，$P_i \in [0, 1]$ と $\Delta R_i \in [-\infty, +\infty]$ は明らかな一対一対応の関係にある：

$$P_i = \frac{1}{1 + \exp[-\beta_i \Delta R_i]} = \mathrm{sig}(\beta_i \Delta R_i) \tag{4.2}$$

ここで，$\mathrm{sig}(x) = 1/(1 + \exp[-x])$ はシグモイド関数である．

均衡解を分析するため，$P_i$ あるいは $\Delta R_i$ の局所的な振る舞いを記述する必要がある．そこで，累積報酬 $R_{i,x}$ の確率過程の平均的挙動が不動となる条件を求める．累積報酬の式 (4.1) は漸化式

$$R_{i,x}(X_{t+1}) = \alpha_i \left[ R_{i,x}(X_t) + \delta(x, x_{i,t})\, f_i(x_t) \right]$$

としても表現でき，その変化量は以下となる：

$$R_{i,x}(X_{t+1}) - R_{i,x}(X_t) = (\alpha_i - 1)\, R_{i,x}(X_t) + \alpha_i\, \delta(x, x_{i,t})\, f_i(x_t)$$

本章の関心は確率差分方程式の均衡解なので，実現値 $\delta(x, x_{i,t})\, f_i(x_t)$ に代わり，その平均的な挙動すなわち期待値 $\gamma_{i,x}$（のちに定義）で置き換える．このとき，ある点 $(R_{1,\mathrm{C}}, R_{1,\mathrm{D}}, R_{2,\mathrm{C}}, R_{2,\mathrm{D}})$ における確率的な振る舞いはその平均的な変化量 $V_{i,x}$ で代表できる：

$$V_{i,x} = (\alpha_i - 1)\ R_{i,x} + \alpha_i \gamma_{i,x}$$

ここで，期待利得 $\gamma_{i,x}$ は以下である．

$$\gamma_{1,x} = \sum_y P_1(x)\, P_2(y)\, f_1(x, y)$$
$$\gamma_{2,y} = \sum_x P_1(x)\, P_2(y)\, f_2(x, y)$$

この変化量 $V_{i,x}$ を $\Delta R_i = R_{i,\mathrm{C}} - R_{i,\mathrm{D}}$ の空間へ対応づければ

$$\begin{aligned}\Delta V_i &= V_{i,\mathrm{C}} - V_{i,\mathrm{D}} \\ &= (\alpha_i - 1)\ (R_{i,\mathrm{C}} - R_{i,\mathrm{D}}) + \alpha_i\ (\gamma_{i,\mathrm{C}} - \gamma_{i,\mathrm{D}}) \\ &= (\alpha_i - 1)\ \Delta R_i + \alpha_i \Delta\gamma_i\end{aligned}$$

となる．変化量 $\Delta V_i = 0$ となるとき，その $(\Delta R_1, \Delta R_2)$ においてこの確率過程は平均的に見れば不動となる．これはこの確率的近似モデルにおける均衡解とみなせ，本来のモデルにおける解の存在を示唆するものと考えられる．この分析方法は各点 $\Delta R_i \in [-\infty, +\infty]$ の変化の向きを調べるので「ベクトル場近似」と呼ぶ．

$\Delta R_i \in [-\infty, +\infty]$ に比べ，有界の $P_i \in [0, 1]$ は扱いやすい．そこで，$\Delta V_i$ を行動確率のペア $(P_1, P_2)$ の関数 $\Upsilon_i(P_1, P_2) := \Delta V_i$ として記述する．まず，$\Delta\gamma_i$ は明らかに $P_1$，$P_2$ の関数であり，これを $\Gamma_i(P_1, P_2) := \Delta\gamma_i$ と明示的に表記する．また，式 (4.2) は逆関数

$$\beta_i \Delta R_i = \log \frac{P_i}{1 - P_i} = \mathrm{logit}(P_i)$$

をもち，代入すれば累積利得の項が消え，$(P_1, P_2)$ の関数をえる：

$$\Upsilon_i(P_1, P_2) = (\alpha_i - 1)\ \frac{1}{\beta_i} \mathrm{logit}(P_i) + \alpha_i \Gamma_i(P_1, P_2)$$

### 4.3.2 固定点，ヌルクライン，安定性

$\Upsilon_i(P_1, P_2)$ は点 $(P_1, P_2)$ での平均的挙動を捉える．$P_i$ に関して

$$\{(P_1, P_2) : \Upsilon_i(P_1, P_2) = 0\}$$

をみたす点の集合を $P_i$ のヌルクラインという．また，すべてのヌルクラインの交点を固定点という：

$$\{(P_1, P_2) : \Upsilon_i(P_1, P_2) = 0 \text{ for all } i\}$$

固定点には安定点, 不安定点, 鞍点という 3 種類があり, $\Upsilon_i$ の微分を調べることで判別できる．ある固定点 $(P_1, P_2)$ がすべての $i$ で $\frac{\partial \Upsilon_i(P_1, P_2)}{\partial P_i} < 0$ をみたすとき, 安定点あるいは均衡解という．他方, すべての $i$ で $\frac{\partial \Upsilon_i(P_1, P_2)}{\partial P_i} > 0$ をみたすとき，不安定点という．ある方向について負の傾きをもつが，別の方向について正の傾きをもつとき，鞍点（サドル）という．

### 4.3.3 $\Gamma_i$ の性質と略記

$\Gamma_i$ と利得行列との関係は $\Gamma_1(1,1) = f_1(\mathrm{CC}), \Gamma_1(1,0) = f_1(\mathrm{CD}), \Gamma_1(0,1) = -f_1(\mathrm{DC})$，$\Gamma_1(0,0) = -f_1(\mathrm{DD})$ である．また，$\Gamma_i$ は $P_1$ と $P_2$ に対して線形である．因人のジレンマでは $\Gamma_1(0.5, y \in [0,1]) < 0$ かつ $\Gamma_2(x \in [0,1], 0.5) < 0$ が成りたつ．

$\Gamma_i(P_1, P_2)$ を整理すると以下をえる．

$$
\begin{aligned}
\Gamma_1(P_1,P_2) &= \Delta\gamma_1 = \gamma_{1,C} - \gamma_{1,D} \\
&= P_1P_2\,[f_1(\mathrm{CC}) - f_1(\mathrm{CD}) + f_1(\mathrm{DC}) - f_1(\mathrm{DD})] \\
&\quad + P_1\,[+f_1(\mathrm{CD}) + f_1(\mathrm{DD})] \\
&\quad + P_2\,[-f_1(\mathrm{DC}) + f_1(\mathrm{DD})] \\
&\quad - f_1(\mathrm{DD}) \\
\Gamma_2(P_1,P_2) &= \Delta\gamma_2 = \gamma_{2,C} - \gamma_{2,D} \\
&= P_1P_2\,[f_2(\mathrm{CC}) - f_2(\mathrm{DC}) + f_2(\mathrm{CD}) - f_2(\mathrm{DD})] \\
&\quad + P_2\,[+f_2(\mathrm{DC}) + f_2(\mathrm{DD})] \\
&\quad + P_1\,[-f_2(\mathrm{CD}) + f_2(\mathrm{DD})] \\
&\quad - f_2(\mathrm{DD})
\end{aligned}
$$

以降では表記を単純化して，

$$
\begin{aligned}
\Gamma_1(P_1,P_2) &= P_1P_2X_1 + P_1Y_1 + P_2Z_1 + A_1 \\
\Gamma_2(P_1,P_2) &= P_1P_2X_2 + P_2Y_2 + P_1Z_2 + A_2
\end{aligned}
$$

とする．因人のジレンマでは $X_i > 0$，$Z_i < 0$，$X_i + Z_i > 0$ が成りたつ．

## 4.4 ベクトル場の可視化

この定式化の利点のひとつとして，確率空間 $(P_1, P_2) \in [0,1] \times [0,1]$ 上のベクトル場を可視化することで，ゲームの確率的ダイナミクスの全体像，ヌルクラインの交点を視覚的に捉えることができる．

行動確率 $P_i \in [0,1]$ 上での変化量は $(P_1, P_2)$ の関数として，

$$\Delta P_i = \mathrm{sig}[\mathrm{logit}(P_i) + \beta_i \Upsilon_i(P_1, P_2)] - P_i$$

と書ける．また，$\Upsilon_i(P_1, P_2) = 0$ を整理すると，

$$P_2 = \frac{(1-\alpha_1)/\beta_1 \operatorname{logit}(P_i) - \alpha_1 (P_1 Y_1 + A_1)}{\alpha_1 (P_1 X_1 + Z_1)}$$
$$P_1 = \frac{(1-\alpha_2)/\beta_2 \operatorname{logit}(P_2) - \alpha_2 (P_2 Y_2 + A_2)}{\alpha_2 (P_2 X_2 + Z_2)}$$

がえられる．それぞれ他方の行動確率のみの関数となっており，これらは $\Upsilon_i(P_1, P_2) = 0$ となるペア $(P_1, P_2)$ を一対一対応で指定する．以上の知見を組み合わせると，各点 $(P_1, P_2) \in [0,1] \times [0,1]$ における変化ベクトル（向きと大きさ）およびヌルクラインを描画できる．

図 4.1 は利得行列 3051 において，$\alpha_1 = \alpha_2$ を $0.0, 0.2, 0.4, 0.6, 0.8, 1.0$ と変えたときのベクトル場の変化を示す（矢印の大きさは適度に縮小してある）．$\alpha_i = 0.0$ では強化学習は C と D を等確率で選択するため，すべてのベクトルは唯一の交点 $(P_1, P_2) = (0.5, 0.5)$ の方向を向く．この交点は安定点である．ヌルクラインは $P_1 = 0.5$（赤線）および $P_2 = 0.5$（緑線）である．$\alpha_i = 0.2$，$\alpha_i = 0.4$ と大きくなるにつれ，唯一の交点が $(0, 0)$ の方向へ移動する．$\alpha_i = 0.6$，$\alpha_i = 0.8$ では，さらに４つの交点が出現する．$(1, 1)$ 付近に安定点，$(0.8, 0.8)$ 付近に不安定点，その他は鞍点である．このうち，$P_i < 0.5$ の安定点は相互裏切 DD に対応し，$P_i > 0.5$ の安定点は相互協調 CC に対応づけられる．$\alpha_i \approx 1.0$ では $(1, 0)$ および $(0, 1)$ 付近にさらに２つの交点が出現する．

図 4.2 は利得行列 1220 において $\alpha_1 = \alpha_2$ を $0.0, 0.2, 0.4, 0.6, 0.8, 1.0$ と変えたときのベクトル場の変化を示す．$\alpha_i \leq 0.8$ までは利得行列 3051 と同じように振る舞うが，$\alpha_i \approx 1.0$ においても，交点は直線 $P_1 = P_2$ 上に３つしか存在しない．

以上の結果は「$\alpha_i = 0.0$ ではランダムに振る舞うが，$\alpha_i \to 1.0$ にともない相互協調の実現確率が高まる」という前章の結果と整合的であると思われる．事実，図 4.1 では $\alpha_i > 0.6$ 付近から相互協調に対応する安定点が出現している．他方，この結果は「相互裏切すなわち $P_i < 0.5$ が唯一の安定点である」という別の近似モデルを用いた先行研究の報告 [31, 55] と矛盾する．この点は 4.9.4 節で論じる．

このように，ベクトル場近似では，有限マルコフ過程のように定常分布を定量的に捉えることはできないが，反面，ゲームの確率的ダイナミクスの全体像を捉えることができる．また，後続の節で見るように，ベクトル場近似は，存在する解の集合，解の存在条件，解の安定性などを解析的に分析できるという利点をもつ．しかしながら，あくまでも近似

であって，本来のモデルがとりえない状態（点）も含めて分析対象としている点を注意する必要がある．

本章の関心は，相互裏切 DD に対応する安定点，相互協調 CC に対応する安定点が，どのような条件で生じるかという問いである．以降ではそれぞれ DD 優位解，CC 優位解と呼ぶ．囚人のジレンマにおいて関心があるのは CC 優位解の存在条件である．後続の節では，まず図 4.2 に対応する「特殊ケース」を分析する．

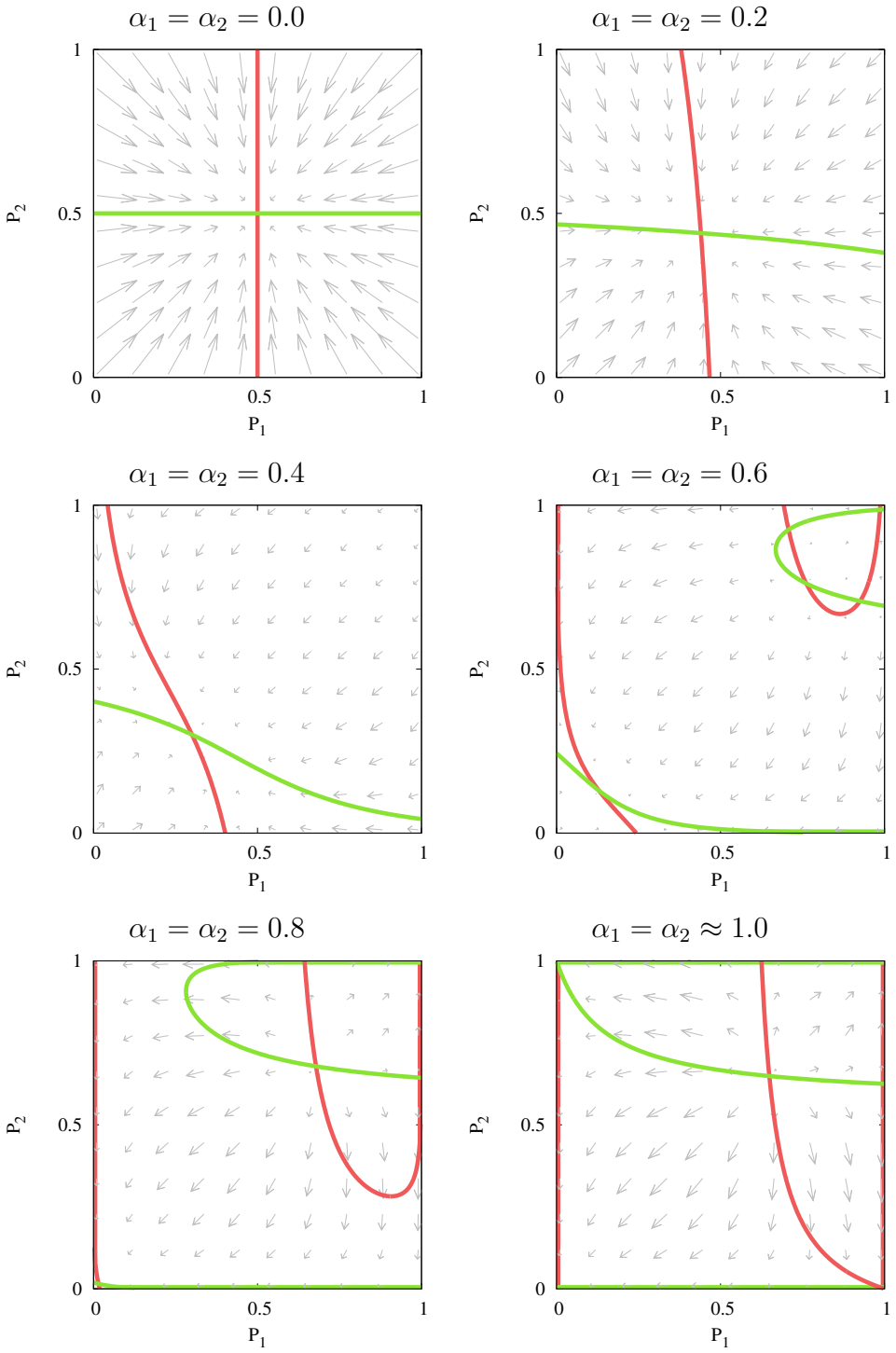



図 4.1: 近似モデルのベクトル場．利得行列 $(3, 0, 5, 1)$

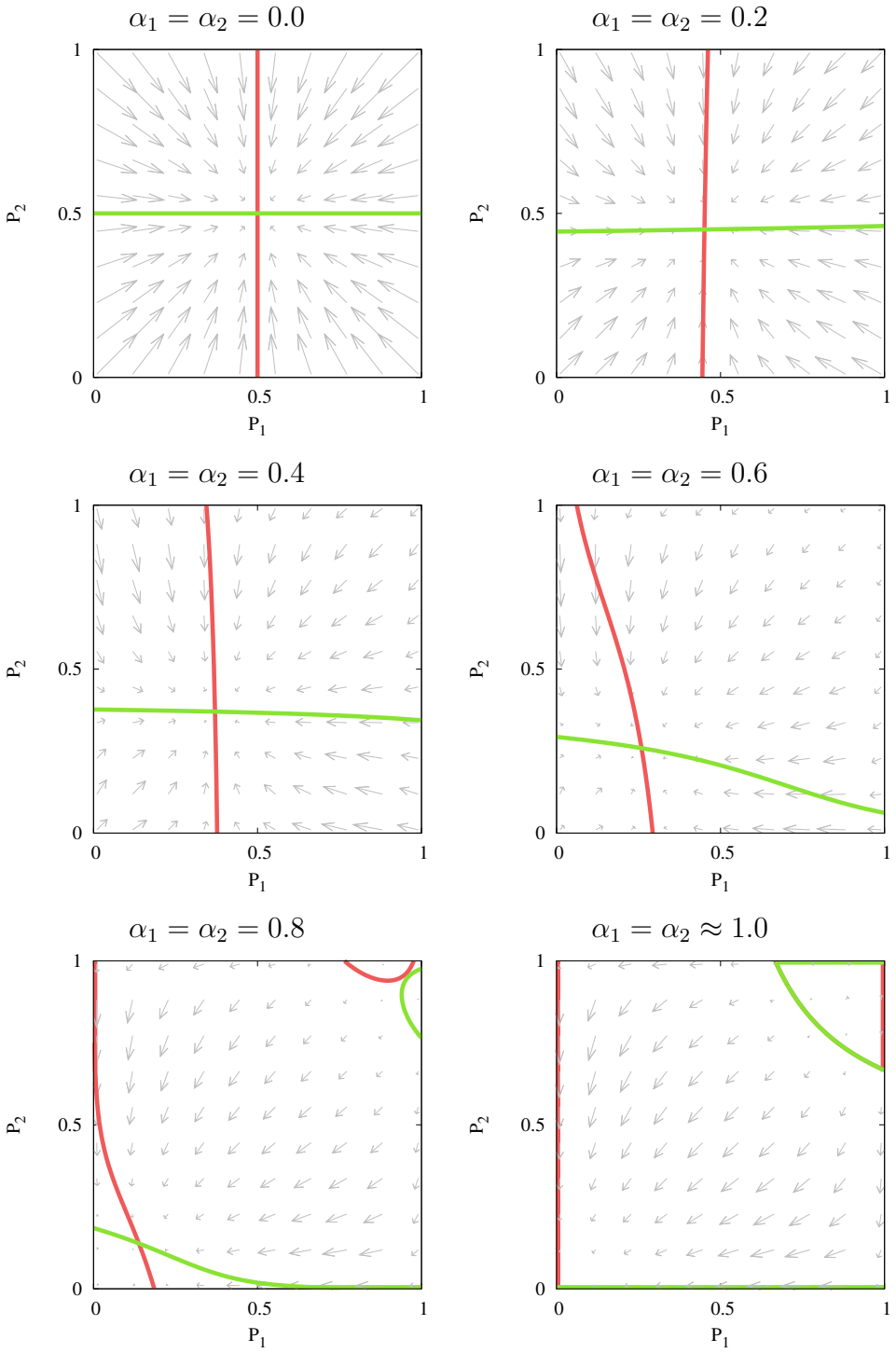



図 4.2: 近似モデルのベクトル場．利得行列 $(1, -2, 2, 0)$

## 4.5 特殊ケース

利得行列が次の条件をみたすとき，「プレイヤ対称」であるという．

$$f_1(\mathrm{CC}) = f_2(\mathrm{CC}) \quad \text{and} \quad f_1(\mathrm{DD}) = f_2(\mathrm{DD})$$
$$\text{and} \quad f_1(\mathrm{CD}) = f_2(\mathrm{DC}) \quad \text{and} \quad f_1(\mathrm{DC}) = f_2(\mathrm{CD})$$

加えて，各プレイヤの利得行列が

$$f_1(\mathrm{CD}) = -f_1(\mathrm{DC}) = f_2(\mathrm{DC}) = -f_2(\mathrm{CD})$$

をみたす場合を考える．これは利得行列 1220（図 4.2）に対応する．このとき，次の補題をえる．

**Lemma 4.5.1.** 初期条件を $P_1 = P_2$，パラメータを $\alpha_1 = \alpha_2$ かつ $\beta_1 = \beta_2$ とする．このとき，本来のモデルおよび近似モデルにおいて常に $P_1 = P_2$ をみたす．

*Proof.* $\Delta R_i(X_{t+1}) = \alpha_i \left[\Delta R_i(X_t) + \delta(\mathrm{C}, x_{i,t}) f_i(x_t) - \delta(\mathrm{D}, x_{i,t}) f_i(x_t)\right]$ である．特殊ケースの仮定の下では，どのような $(x_1, x_2) \in \{\mathrm{CC}, \mathrm{CD}, \mathrm{DC}, \mathrm{DD}\}$ に対しても $[\delta(\mathrm{C}, x_1) - \delta(\mathrm{D}, x_1)] f_1(x_1, x_2) = [\delta(\mathrm{C}, x_2) - \delta(\mathrm{D}, x_2)] f_2(x_1, x_2)$ をみたす．したがって，次時点でも $\Delta R_1(X_{t+1}) = \Delta R_2(X_{t+1})$ をみたし，ゆえに $P_1(\mathrm{C}|X_{t+1}) = P_2(\mathrm{C}|X_{t+1})$ をみたす． □

この特殊ケースは囚人のジレンマの主要な解を含みながらも，$P_1 = P_2$ の直線上のみを考えればよく，数理的に扱いやすく，また視覚的にも捉えやすい．ここでは，$P_1 = P_2$ かつ $\Gamma_1 = \Gamma_2$ であるため，これらを $P := P_i$ および $\Gamma(P) := \Gamma_i(P, P)$ と記す．また，$X_1 = X_2$，$Y_1 = Y_2 = Z_1 = Z_2$，$A_1 = A_2$ であるため，これらも添字 $i$ を省略し，$Y_i$ の代わりに $Z_i$ を用いる．以降では，不要なときには，すべての変数について添字 $i$ を省略する．囚人のジレンマでは $X > 0$ だから，$\Gamma(P)$ は下方向の凸の二次関数である．その頂点は $P' = -(Y + Z)/(2X) = -Z/X$ で，$f(\mathrm{CD}) = -f(\mathrm{DC})$ をみたす囚人のジレンマでは $\Gamma(P') = -Z^2/X + A < 0$ である．

### 4.5.1 特殊ケースの可視化

固定点の条件 $\Delta V = \Upsilon(P) = 0$ より，

$$\begin{aligned}
&\Upsilon(P) = (\alpha - 1)/\beta \ \mathrm{logit}(P) + \alpha \ \Gamma(P) = 0 \\
&\iff (1-\alpha) \ \alpha/\beta \ \mathrm{logit}(P) = \Gamma(P) \\
&\iff (1-\alpha) \ \alpha/\beta = \Gamma(P)/\mathrm{logit}(P)
\end{aligned}$$

などの等式がえられる．ここで，$0 \leq \alpha \leq 1$ より，$\phi_\alpha := (1-\alpha)\,\alpha/\beta \geq 0$ である．図 4.3 に，利得行列 1220 とその変種について，それぞれの等式を図示した（$\beta = 1$）．これら 2 つの関数の交点がゲームの均衡解でありうる．左列の図は $\Upsilon(P)$ であり，点 $P$ の安定性を示す．中央の図は $\Gamma(P)$ と $\phi_\alpha \,\mathrm{logit}(P)$ であり，双方の関数の形状がもっとも単純であるため，以降で解の個数の証明に用いる．右列の図はパラメータ $\alpha$ の変化にともなう解の個数の分岐を示す（$\alpha \to 1$ において $\phi_\alpha \to 0$）．

左列および中央の図から，解の個数と各解の安定性が読みとれる．図中 (a) は 1 つの交点をもつが，その周辺で $\Upsilon(P) > 0$ から $\Upsilon(P) < 0$ へ変化しているので安定点である．この唯一の交点は $P < 0.5$ に位置する DD 優位解である．図 4.3 (b), (c) は 3 つの交点をもつが（$\mathrm{logit}(P) \in [-\infty, +\infty]$），中央の交点はその周辺で $\Upsilon(P) < 0$ から $\Upsilon(P) > 0$ へ変化しており不安定点である．左端の交点は $P < 0.5$ に位置する DD 優位解であるが，他方，右端の交点は $P > 0.5$ に位置する CC 優位解である．また，右列の図から，(a) では $\alpha \to 1$ においても CC 優位解は出現しないが，(b), (c) では出現することがわかる．これらの図に言及しながら，以上のことを証明する．

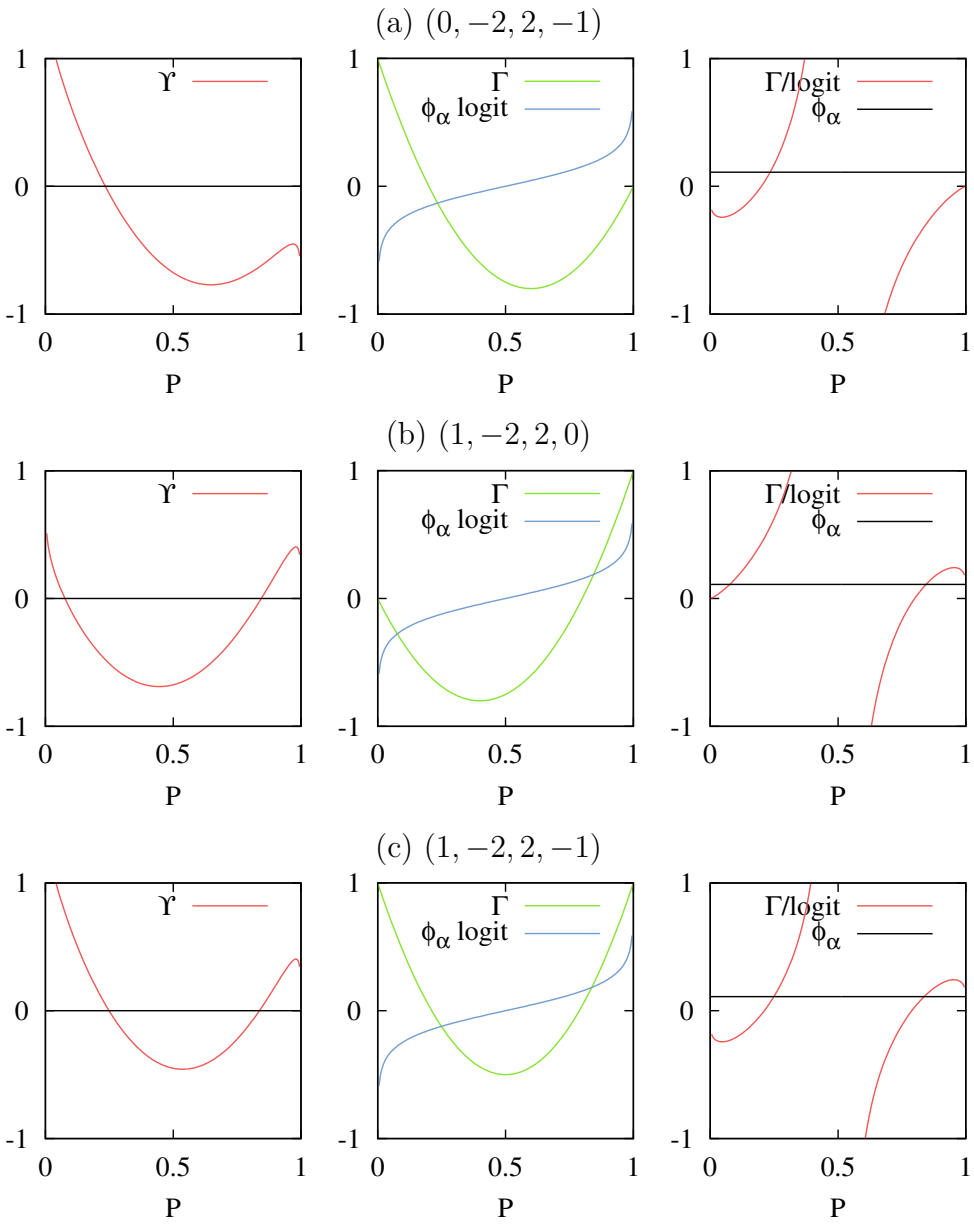



図 4.3: 特殊ケース $P_1 = P_2$ の分析（3 つの異なる利得行列）

### 4.5.2 特殊ケースの DD 優位解

**Lemma 4.5.2.** 囚人のジレンマでただ 1 つ常に存在する安定点は $P < 0.5$ の範囲に存在する DD 優位解である.

*Proof.* DD 優位解が常に存在することを示す．DD 優位解とは $\phi_\alpha \operatorname{logit}(P) = \Gamma(P)$ をみたす安定点 $P < 0.5$ をいう（中央の図）．まず，$\phi_\alpha \operatorname{logit}$ は単調増加関数であり，$\operatorname{logit}(0) = -\infty$，$\operatorname{logit}(0.5) = 0$ である．また，$\Gamma(P)$ は下に凸の二次関数であり，$\Gamma(0) = f(\mathrm{DD})$，$\Gamma(0.5) < 0$ である．したがって，中間値の定理より交点をもつ．$\Upsilon(P) > 0 \iff \phi_\alpha \operatorname{logit}(P) < \Gamma(P)$ を $P = 0$ で常にみたす．それゆえに，もっとも $P = 0$ に近い交点は安定点である． □

### 4.5.3 特殊ケースの CC 優位解

**Lemma 4.5.3.** 囚人のジレンマで $\alpha \to 1$ において CC 優位解の存在する必要条件は $f(\mathrm{CC}) > 0$ である.

*Proof.* $\alpha \to 1$ において CC 優位解の存在する必要条件を示す．CC 優位解とは $\phi_\alpha \operatorname{logit}(P) = \Gamma(P)$ をみたす安定点 $P > 0.5$ をいう（中央の図）．まず，$\phi_\alpha \operatorname{logit}$ は単調増加関数であり，$P > 0.5$ の範囲で $\operatorname{logit}(P) > 0$ である．極端な場合として $\alpha \approx 1$ では，$\phi_\alpha \approx 0$ ゆえに $0.5 < P < 1$ の範囲で $\phi_\alpha \operatorname{logit}(P) \approx 0$ となるが，$\phi_\alpha \operatorname{logit}(1) = +\infty$ をみたす（$\alpha < 1$）．したがって，中間値の定理より $\alpha \to 1$ において $\Gamma$ と $\phi_\alpha \operatorname{logit}$ が交点をもつ必要条件は $\Gamma(1) = f(\mathrm{CC}) > 0$ である．$\Upsilon(P) < 0 \iff \phi_\alpha \operatorname{logit}(P) > \Gamma(P)$ を $P = 1$ で常にみたす．それゆえに，もっとも $P = 1$ に近い交点は安定点である． □

### 4.5.4 特殊ケースの解の個数と安定性

$P_1 = P_2$ をみたす特殊ケースにおいて，解が少なくとも 1 つ，多くとも 3 つであることを証明する．証明には補助的に図 4.4 に言及する．

ある点 $P^*$ で logit と $\Gamma$ が接するとき，次の条件をみたす.

$$\operatorname{logit}(P^*) = \Gamma(P^*) \quad \text{and} \quad \tfrac{\partial}{\partial P} \operatorname{logit}(P^*) = \tfrac{\partial}{\partial P} \Gamma(P^*)$$

logit と $\Gamma$ は図 4.4 に示す 2 通りの仕方で接点をもちうる．この接点をそれぞれ $P^*$，$P^{**}$（ただし $P^* < P^{**}$）とし，また，各点で接する関数を $\Gamma^*$，$\Gamma^{**}$ とする．

*Proof.* 少なくとも 1 つの交点をもつことを示す．$P \in [0,1]$ 上に有限の値をもつ任意の関数 $\Gamma$ について，$f(P) = \mathrm{logit}(P) - \Gamma(P)$ は $f(0) = -\infty$ かつ $f(1) = +\infty$ をとる連続関数である．中間値の定理より，$f(P) = 0$ をみたす $P$ が存在する． □

任意の $\Gamma$ が 3 つの交点（接点を含む）をもつとき，次の条件をみたす．

$$\Gamma(P^*) < \Gamma^*(P^*) \quad \text{and} \quad \Gamma(P^{**}) > \Gamma^{**}(P^{**})$$

*Proof.* この条件をみたすとき，$0 \leq P^* < P^{**} \leq 1$ の各区間において，それぞれ 1 つの交点が存在することを示す．まず，$\Gamma(P^*) < \Gamma^*(P^*)$ を仮定する．$P \in [0, P^*]$ の範囲で $f(0) = -\infty$ かつ $f(P^*) > 0$ だから，中間値の定理により，$f(P) = 0$ をみたす $P$ が存在する．次に，$\Gamma(P^{**}) > \Gamma^{**}(P^{**})$ を仮定する．$P \in [P^{**}, 1]$ の範囲で $f(P^{**}) < 0$ かつ $f(1) = +\infty$ だから，中間値の定理により，$f(P) = 0$ をみたす $P$ が存在する．上記 2 つの仮定をみたすとき，$P \in [P^*, P^{**}]$ の範囲で $f(P^*) > 0$ かつ $f(P^{**}) < 0$ だから，中間値の定理により，$f(P)$ をみたす $P$ が存在する． □

任意の $\Gamma$ が 1 つの交点をもつとき，次の条件をみたす．

$$\begin{aligned} \text{either} \quad & \Gamma(P^*) > \Gamma^*(P^*) \\ \text{or} \quad & \Gamma(P^{**}) < \Gamma^{**}(P^{**}) \end{aligned}$$

任意の $\Gamma$ が 2 つの交点（1 つは接点）をもつとき，次の条件をみたす．

$$\begin{aligned} \text{either} \quad & \Gamma(P^*) = \Gamma^*(P^*) \\ \text{or} \quad & \Gamma(P^{**}) = \Gamma^{**}(P^{**}) \end{aligned}$$

*Proof.* 次に，4 つ以上の交点をもたないことを示す．$\Gamma$ の頂点を与える $P$ を $x$ とする．logit は単調増加（勾配 $> 0$）であるから，$P^* > x$ かつ $P^{**} > x$ をみたすが，$P < P^*$ をみたす交点は $P' \leq x$ か $P' > x$ でありうる．まず，$P' \leq x$ のとき，$P \in [0, x]$ の範囲で $\Gamma$ は単調減少だから，$P'$ を除いて交点は存在せず，$P \in [P^*, P^{**}]$，$P \in [P^{**}, 1]$ それぞれの区間で多くとも 1 つの交点をもちうる．他方，$P' > x$ のとき，$P \in [x, 1]$ の範囲で $\Gamma$ は単調増加だから，$P \in [x, P^*]$，$P \in [P^*, P^{**}]$，$P \in [P^{**}, 1]$

のそれぞれの区間で多くとも 1 つの交点をもちうる．よって，4 つ以上の交点をもたない． □

以上の証明から，logit と $\Gamma$ は (a) $P \in [0, P^*)$ に 1 つの交点をもつ；(a') $P \in [0, P^*)$ に 1 つの交点，$P = P^{**}$ に 1 つの接点をもつ；(b) $P \in [0, P^*)$，$P \in (P^*, P^{**})$，$P \in (P^{**}, 1]$ にそれぞれ 1 つの交点をもつ；(c') $P \in (P^{**}, 1]$ に 1 つの交点，$P = P^*$ に 1 つの接点をもつ；(c) $P \in (P^{**}, 1]$ に 1 つの交点をもつ，のいずれかとなる．したがって，

**Lemma 4.5.4.** $P_1 = P_2$ をみたす特殊ケースにおいて，解は少なくとも１つ，多くとも３つある．

logit の形状から，(a)，(c) では，ただ 1 つの交点は安定点である．また (b) では，両外側に安定点を 1 つずつ，内側に不安定点を 1 つもつ．(a')，(c') では，$P^{**}$ で接するか，$P^*$ で接するかで異なるが，その接点は鞍点（不安定），他の交点は安定点である．表 4.1 にこの結果を要約した．

表 4.1: 安定点（●），不安定点（○），その周辺でのダイナミクス．中央の 2 つの矢印は $P^*$，$P^{**}$ 上での動きを表す（鞍点はその近辺の動き）

| | 交点数 | 接点 | ダイナミクス |
|---|---|---|---|
| a | 1 | | $\rightarrow \bullet \leftarrow \quad \leftarrow \quad \leftarrow$ |
| a' | 2 | $P^{**}$ | $\rightarrow \bullet \leftarrow \quad \leftarrow \circ \leftarrow$ |
| b | 3 | | $\rightarrow \bullet \leftarrow \circ \rightarrow \bullet \leftarrow$ |
| c' | 2 | $P^*$ | $\rightarrow \circ \rightarrow \quad \rightarrow \bullet \leftarrow$ |
| c | 1 | | $\rightarrow \quad \rightarrow \quad \rightarrow \bullet \leftarrow$ |

### 囚人のジレンマでは CC 優位解だけの場合はない

**Lemma 4.5.5.** 囚人のジレンマでは (c) や (c') となる解は存在しない．

*Proof.* 囚人のジレンマでは $\Gamma(0.5) < 0$ をみたす．$\mathrm{logit}(0.5) = 0$ であるから，$P \in [0, 0.5)$ の区間に $\Gamma(P) < \mathrm{logit}(P)$ をみたす $P$ が必ず存在する．(c) および (c') となる解が存在するには $\Gamma(P^*) \geq \Gamma^*(P^*)$ でなければならないが，これはすべての $P \in [0, P^*]$ において $\Gamma(P) > \mathrm{logit}(P)$ を意味し，矛盾する． □

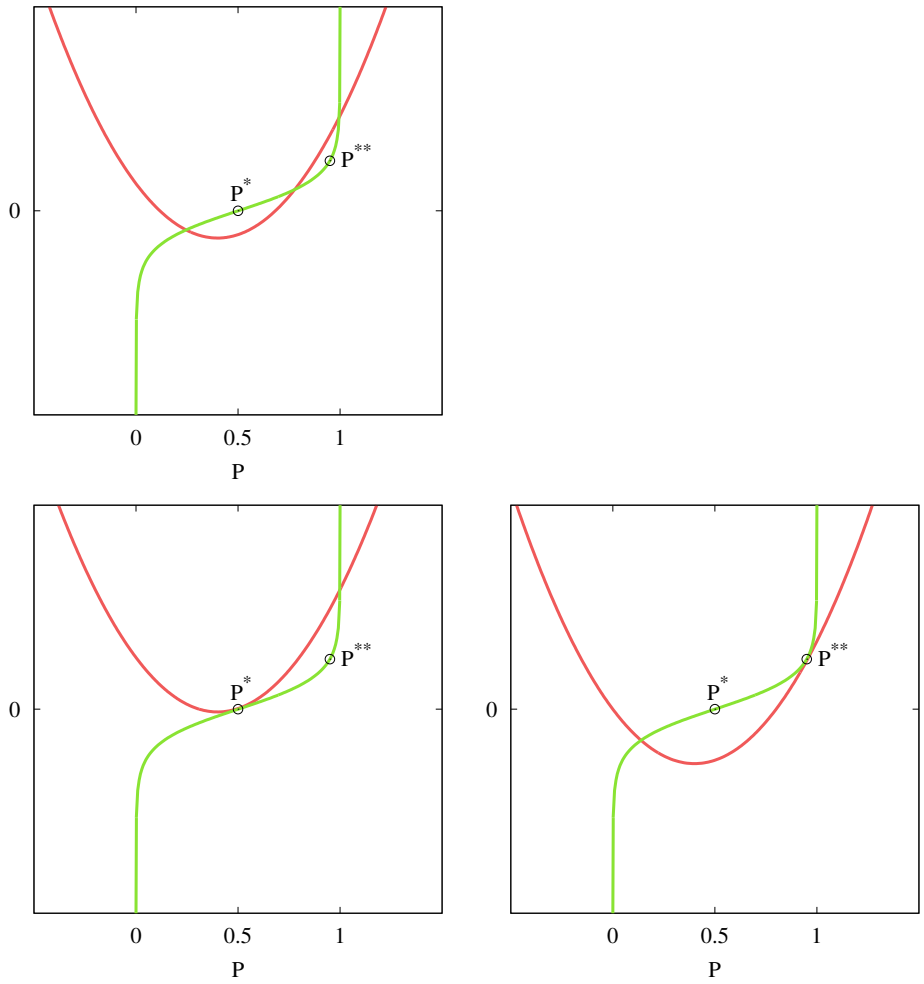



図 4.4: 特殊ケース $P_1 = P_2$ における logit と $\Gamma$ の交点と接点

## 4.6 一般ケース

利得行列の対称性を仮定しない一般の場合 $P_1 \neq P_2$ を考える．このとき，$\Gamma_1$ と $\Gamma_2$ を区別する必要があるが，まったく同様の結果を $\Gamma_2$ に対してえられるため，$\Gamma_1$ のみを考える．$\Gamma_1$ は $P_1$ と $P_2$ の多重線形関数である．logit が $P_1$ のみの関数であるため，$P_2$ を任意の定数とみなすことで前節と同様の証明を展開できる．証明には補助的に図 4.5 に言及する．

### 4.6.1 一般ケースの DD 優位解

**Lemma 4.6.1.** 囚人のジレンマでただ 1 つ常に存在する安定点は $P_1 < 0.5$ かつ $P_2 < 0.5$ の範囲に存在する DD 優位解である．

*Proof.* $\Gamma_1(P_1, P_2)$ は $P_1$ の線形関数であり，その傾きの符号は $P_2$ に依存する．囚人のジレンマでは，$P_2$ に依らず，$\Gamma_1$ と $\phi_{\alpha_1}\,\mathrm{logit}$ が $P_1 < 0.5$ の範囲で交点をもつことを示す．まず，$P_1 \in [0, 0.5]$ の範囲で $\mathrm{logit}(P_1) \in [-\infty, 0]$ である．$\Gamma_1$ は $\Gamma_1(0.5, P_2) < 0$ をみたす直線であるから，傾きの符号に関わらずどのような $P_2$ に対しても， $P_1 < 0.5$ の範囲で単調増加関数 logit と唯一の交点をもつ．$\Gamma_1(0, P_2) > \phi_{\alpha_1}\,\mathrm{logit}(0)$ であるから，その交点は安定点である．$\Gamma_2$ についても同様である．したがって，$\Gamma_1$ は任意の $P_2 < 0.5$，他方，$\Gamma_2$ は任意の $P_1 < 0.5$ に関してヌルクラインをもつから，これらのヌルクラインは交点をもち，その交点は安定点である． □

### 4.6.2 一般ケースの CC 優位解

**Lemma 4.6.2.** 囚人のジレンマで $\alpha_1 \to 1$ かつ $\alpha_2 \to 1$ において CC 優位解の存在する必要条件は $f_1(\mathrm{CC}) > 0$ かつ $f_2(\mathrm{CC}) > 0$ である．

*Proof.* 特殊ケースと同様に極端な場合 $\alpha_i \approx 1$ を考える．$\alpha_i \approx 1$ では $\phi_{\alpha_i} \approx 0$ ゆえに $0.5 < P_i < 1$ の範囲で $\phi_{\alpha_i}\,\mathrm{logit}(P_i) \approx 0$ となるが，$\phi_{\alpha_i}\,\mathrm{logit}(1) = +\infty$ をみたす($\alpha_i < 1$)．ゆえに，$\alpha_1 \approx 1$ において $\Gamma_1$ と logit が交点をもつための条件は $\Gamma_1(1, 1) = f_1(\mathrm{CC}) > 0$ である．$\Gamma_1(1, P_2) < \phi_{\alpha_i}\,\mathrm{logit}(1)$ であるから，その交点は安定点である．$\Gamma_2$ においても同様である．したがって，$\alpha_1 \approx 1$，$\alpha_2 \approx 1$ の条件のもと，$P_1 \approx P_2 \approx 1$ 付近において $\Gamma_1$ と $\Gamma_2$ のヌルクラインが交点をもつための必要条件は

$$f_1(\mathrm{CC}) > 0 \quad \text{and} \quad f_2(\mathrm{CC}) > 0$$

であり，その交点は安定点である． □

### 4.6.3 一般ケースの解の個数と安定性

ある点 $P_1^*$ で接する条件は，任意の $P_2$ を用いて，

$$\mathrm{logit}(P_1^*) = \Gamma(P_1^*) \quad \text{and} \quad \tfrac{\partial}{\partial P_1}\mathrm{logit}(P_1^*) = \tfrac{\partial}{\partial P_1}\Gamma(P_1^*, P_2)$$

であり，図 4.5 に示す 2 通りの仕方で接点をもちうる．この接点をそれぞれ $P_1^*$，$P_1^{**}$（ただし $P_1^* < P_1^{**}$）とし，また，各点で接する関数を $\Gamma^*$，$\Gamma^{**}$ とする．

必ず 1 つの交点をもつこと，および 3 つの交点（鞍点を含む）をもつことは前節と同様に証明できる．1 つの交点をもつ条件，2 つの交点をもつ条件も同様である．ただし，3 つの交点をもつ条件は，任意の $P_2$ を用いて，次式である．

$$\Gamma(P_1^*) < \Gamma^*(P_1^*, P_2) \quad \text{and} \quad \Gamma(P_1^{**}) > \Gamma^{**}(P_1^{**}, P_2)$$

4 つ以上の交点をもたないことも前節と同様に証明できる．

*Proof.* 4 つ以上の交点をもたないことを示す．$\Gamma$ が $P_1$ に対して単調増加か単調減少かは利得行列と $P_2$ に依存する．単調減少の場合，logit が単調増加だから，ただ 1 つの交点しかもたない．単調増加の場合，logit も単調増加だから，$0 \le P_1^* < P_1^{**} \le 1$ の各区間で多くとも 3 つの交点をもちうる． □

以上から，任意の $P_2$ を考えることで，前節と同様の結果（表 4.1）をえられた．$\Gamma_2$ に対しても任意の $P_1$ を考えれば同様の結果をえられる．

$\Gamma_1$ と $\Gamma_2$ は独立であるから，各交点集合の直積がゲームの解の集合となり，多くとも $3 \times 3$ の解が存在しうる．また，すべての $P_2 \in [0, 1]$ で $\Gamma_1(0.5, P_2) < 0$ であるため，4.5.4 節の議論は成り立つ．4.5.4 節の結果より，少なくとも 1 つの DD 優位解が常に存在する．したがって，

**Lemma 4.6.3.** 一般ケース $P_1 \neq P_2$ において，解は少なくとも 1 つ，多くとも 9 つある．

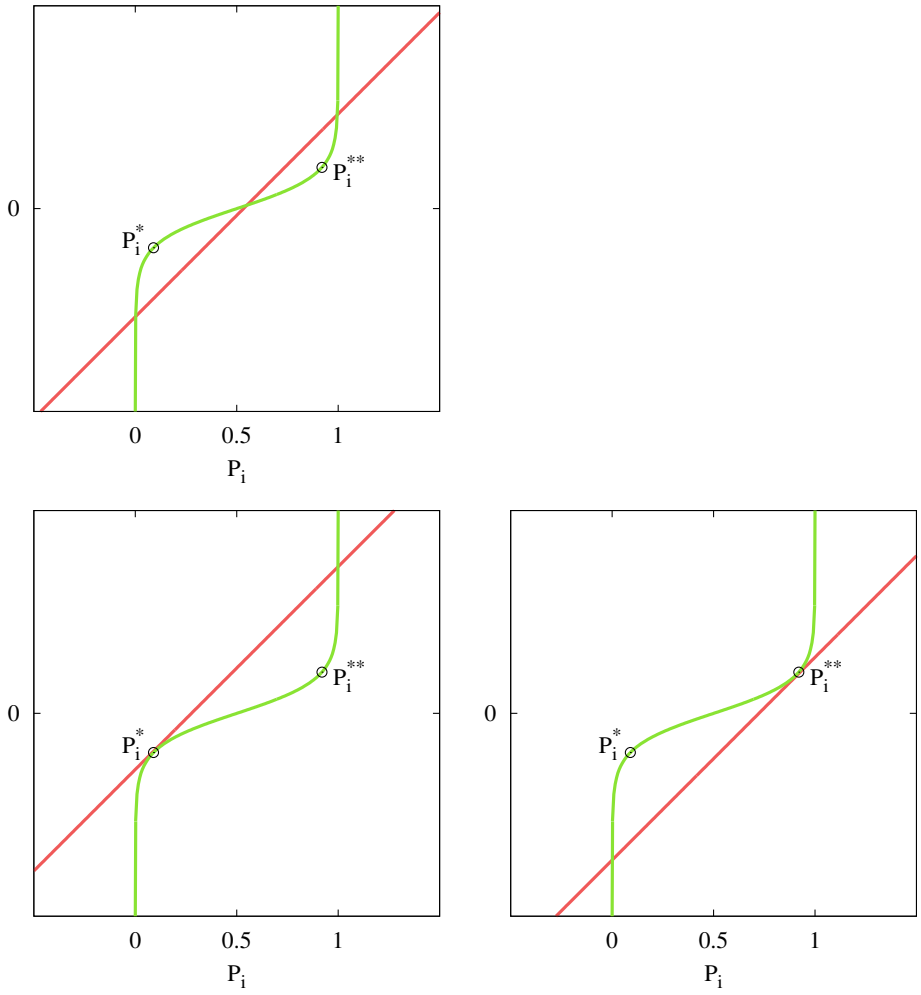



図 4.5: 特殊ケース $P_1 \neq P_2$ における logit と $\Gamma$ の交点と接点

### 一般ケースの解の安定性

これまでの知見を踏まえて，$\Gamma_1$ と $\Gamma_2$ を同時に考え，2 人ゲームの全体像を捉える．このとき，各接点の直積集合 $\{P_1^*, P_1^{**}\} \times \{P_2^*, P_2^{**}\}$ が分類の基点となる．$\Gamma_i$，logit の構造から，表 4.1 に示す 5 つのクラスが実現しうるので，$5 \times 5 = 25$ 通りの交点がありえる．各交点の各軸方向への安定性を，安定点 •，不安定点 ∘，（不安定）鞍点 “ ” としてペア $P_1/P_2$ と表現すれば以下をえる．

$$
\begin{array}{c}
 \\
P_2^{**} \\
 \\
P_2^* \\
 
\end{array}
\begin{array}{c}
\begin{array}{ccccc} P_1^* & & & P_1^{**} & \end{array} \\
\left(\begin{array}{ccccc}
\bullet/\bullet & /\bullet & \circ/\bullet & /\bullet & \bullet/\bullet \\
\bullet/ & / & \circ/ & / & \bullet/ \\
\bullet/\circ & /\circ & \circ/\circ & /\circ & \bullet/\circ \\
\bullet/ & / & \circ/ & / & \bullet/ \\
\bullet/\bullet & /\bullet & \circ/\bullet & /\bullet & \bullet/\bullet
\end{array}\right)
\end{array}
$$

ここで，•/• は全軸方向への安定点，∘/∘ は全軸方向への不安定点を示す．

ベクトル場を可視化して，上記の結果を確認する．図 4.6 は，$\alpha_i \approx 1.0$ において，すべて正の値をもつ利得行列 $(7, 4, 9, 5)$ のベクトル場である．この場合，$f_1(\mathrm{CC}) = 7 > 0$ なので CC 優位解をもち，また $f_1(\mathrm{CD}) = 4 > 0$ なので $\Gamma_1(0, 1)$ 上にも解をもつ．これに対して，図 4.6 はすべて負の値をもつ利得行列 $(-3, -6, -1, -5)$ のベクトル場である．この場合，$f_1(\mathrm{CC}) = -3 \leq 0$ なので，CC 優位解は存在しない．また，唯一の DD 優位解はランダムな振る舞いを示す $(0.5, 0.5)$ に近接している．

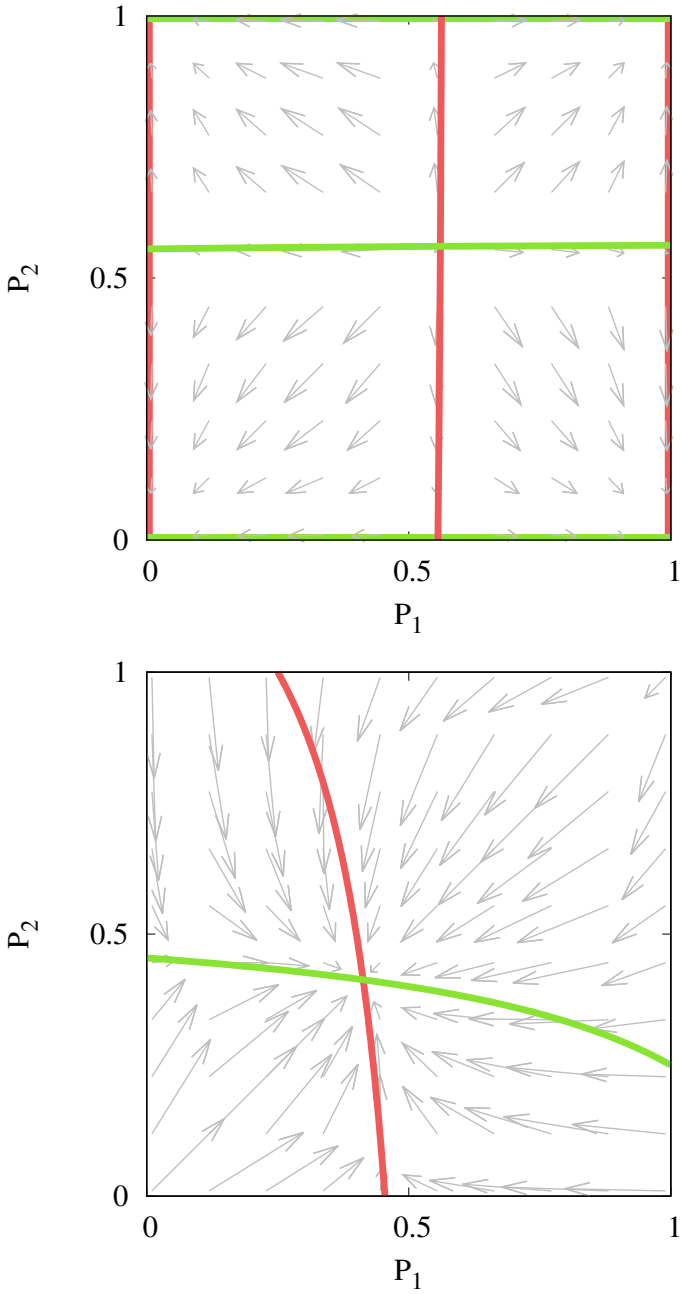



図 4.6: 上図：利得行列 $(7, 4, 9, 5)$，下図：利得行列 $(-3, -6, -1, -5)$. いずれも囚人のジレンマの条件をみたす

## 4.7　CC 優位解へ到達しやすい利得行列

上記の知見を踏まえて，CC 優位解へ到達しやすい利得行列を導出する．ここでは利得行列の値に関心があるので $\alpha_i \approx 1$ を仮定する．

ある $P_2$ を所与として，$\Gamma_1(P_1, P_2)$ は線形である．$\alpha_1 \approx 1$ を仮定すると，logit は $P_1 \in (0,1)$ で $\mathrm{logit}(P_1) \approx 0$ と近似できる．図 4.7 左図はこれを示す．図中，黒丸は $\Gamma_1(0.5, P_2)$ である．DD 優位解，CC 優位解の実現されやすさは (a)，(b) で示される $\Gamma_1$ と logit の間の面積と対応づけられ，(a) の面積を最小化し，(b) の面積を最大化することになる．囚人のジレンマでは $\Gamma_1(0.5, P_2) < 0$ なので，(b) の面積が (a) の面積を上回ることはない．特殊ケース $P_1 = P_2$ として見れば，図 4.7 右図のように，(a) の面積を最小化し，(b) の面積を最大化することに対応する．

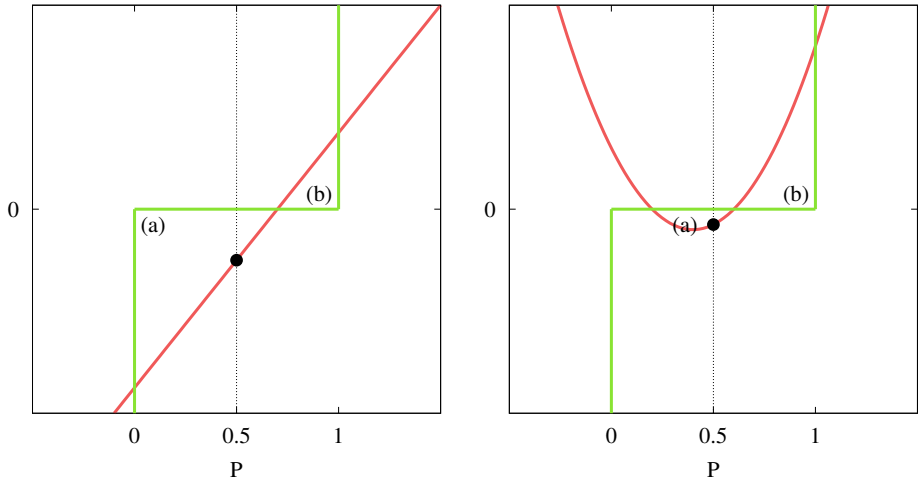



図 4.7: $\Gamma_1(P_1, P_2)$（$P_2$ を定数とみなす）と $\phi_{\alpha_1}\,\mathrm{logit}(P_1)$（ただし $\alpha_i \approx 1$）．左図：一般ケース，右図：特殊ケース．

左図から，$\Gamma_i$ の傾き $\frac{\partial \Gamma_i}{\partial P_i}$ が正のときは $P_i > 0.5$ に解をもち，その傾きが正方向に大きいほど (a) の面積とともに (b) の面積も大きくなり，解は $P_i = 1$ と $P_i = 0$ に近づく．他方，$\Gamma_i$ の傾きが負のときは $P_i < 0.5$ にしか解をもたず，その傾きが負方向に大きいほど (a) の面積は小さくなり，解は $P_i = 0.5$ に近づく．また黒丸に関しては，その値が 0 に近いほど (a) の面積は小さくなり，反対に (b) の面積は大きくなる．

これらの性質から，DD 優位解は $(P_1, P_2) \in [0, 0.5) \times [0, 0.5)$ にしか存在しないので，この範囲では $\Gamma_i$ の傾きを負にすることで，DD 優位解

はランダムに近い位置に出現しやすい．また，CC 優位解は $(P_1, P_2) \in (0.5, 1] \times (0.5, 1]$ にしか存在しないので，この範囲では $\Gamma_i$ の傾きを正にすることで，CC 優位解へと吸引される領域は大きくなる．したがって，その中間点 $P_i = 0.5$ では傾きを 0 にすることが望ましい．明らかに，プレイヤ対称の場合ほど，CC 優位解へと向かう領域は大きくなる．

その制約条件を導く．まず，$\Gamma_1$ の各方向への傾きが 0 となるのは

$$\frac{\partial}{\partial P_1}\Gamma_1(P_1, P_2) = X_1 P_2 + Y_1 = 0$$
$$\frac{\partial}{\partial P_2}\Gamma_1(P_1, P_2) = X_1 P_1 + Z_1 = 0$$

だから，$P_1 = -Z_1/X_1$，$P_2 = -Y_1/X_1$ をえる．この点が直線 $P_1 = P_2$ 上にあるとすれば，$Y_1 = Z_1$，すなわち

$$f_1(\mathrm{CD}) = -f_1(\mathrm{DC}) \tag{4.3}$$

をみたす．このとき，$\Gamma_1(-Z_1/X_1, -Y_1/X_1) = \Gamma_1(0.5, 0.5)$ であるなら

$$[f_1(\mathrm{CC}) + f_1(\mathrm{DD})]^2 = [f_1(\mathrm{CD}) + f_1(\mathrm{DC})]^2$$

をみたす．式 (4.3) とあわせて，制約条件

$$f_1(\mathrm{CC}) = -f_1(\mathrm{DD}) \quad \text{and} \quad f_1(\mathrm{CD}) = -f_1(\mathrm{DC})$$

をえる．この制約条件から，$\Gamma_1(0.5, 0.5) = [f_1(\mathrm{CC}) - f_1(\mathrm{DC})]/2 < 0$ となり，$f_1(\mathrm{CC}) - f_1(\mathrm{DC})$ を 0 に近づければ $\Gamma_1(0.5, 0.5) < 0$ もまた 0 に近づくことがわかる．

利得行列に関して残る不明な点は $f_1(\mathrm{CC}) = -f_1(\mathrm{DD})$ の大きさである．$\Gamma_1(1, 1) = f_1(\mathrm{CC})$ であるから，$f_1(\mathrm{CC})$ が正の値で大きいほど CC 優位解へ向かうベクトルは大きくなる．他方，$\Gamma_1(0, 0) = -f_1(\mathrm{DD})$ であるから，$-f_1(\mathrm{DD})$ が負の値で大きいほど，DD 優位解から離れるベクトルは大きくなる．このとき，DD 優位解は $(P_1, P_2) = (0.5, 0.5)$ に近接している．したがって，$f_1(\mathrm{CC}) = -f_1(\mathrm{DD})$ は十分に大きくできる．

図 4.8 は利得行列 $(10, -10.1, 10.1, -10)$ のベクトル場を示す．この利得行列は特殊ケースの条件をみたし，解は $P_1 = P_2$ 上に 3 つある．DD 優位解と不安定解は中心付近に存在し，どちらもほぼランダムに行動を選択する．CC 優位解は $P_1 = P_2 \approx 1.0$ 付近に存在する．ベクトル場の全体像はその中心 $(0.5, 0.5)$ を経由して CC 優位解へ向かうと解釈できる．この利得行列に対して，有限マルコフ過程（3 章）による分析を行った結

果を図 4.9 に示す．図から，確かにこの利得行列を用いた場合，相互協調が実現されやすくなっていることがわかる．着目すべきことに，より小さな記憶保持率パラメータ $\alpha_i$ であっても相互協調が実現可能になっている．

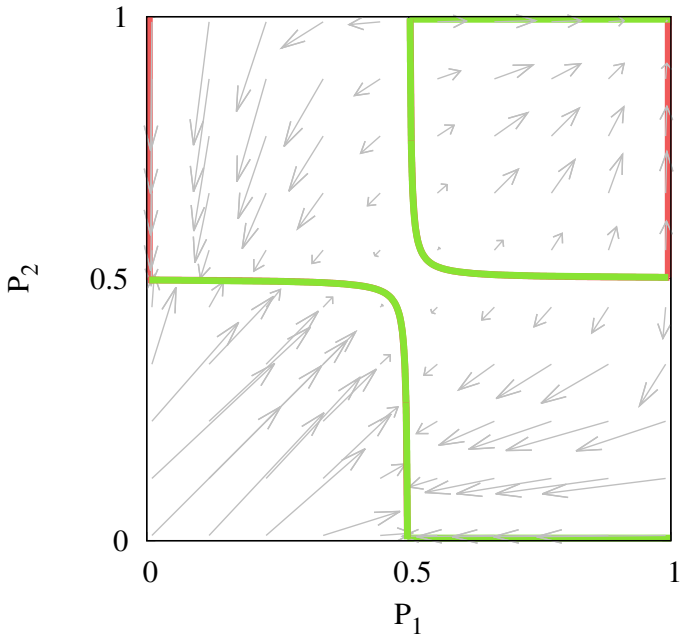



図 4.8: CC 優位解へ到達しやすい利得行列 $(10, -10.1, 10.1, -10)$

## 4.8 無条件報酬の影響

これまで，選択しなかった行動に対する利得（無条件報酬）を 0 と仮定してきた．本節ではこの仮定を置かない場合を論じる．

選択しなかった行動に対する利得を $f_i(\varnothing)$ と記す．簡単のため，$P_{\mathrm{XY}} := P_1(\mathrm{X})P_2(\mathrm{Y})$ と記す．このとき，各行動に対する期待利得 $\gamma_{i,a}$ は

$$\gamma_{1,\mathrm{C}} = P_{\mathrm{CC}}\, f_1(\mathrm{CC}) + P_{\mathrm{CD}}\, f_1(\mathrm{CD}) + P_{\mathrm{DC}}\, f_1(\varnothing) + P_{\mathrm{DD}}\, f_1(\varnothing)$$
$$\gamma_{1,\mathrm{D}} = P_{\mathrm{DC}}\, f_1(\mathrm{DC}) + P_{\mathrm{DD}}\, f_1(\mathrm{DD}) + P_{\mathrm{CC}}\, f_1(\varnothing) + P_{\mathrm{CD}}\, f_1(\varnothing)$$

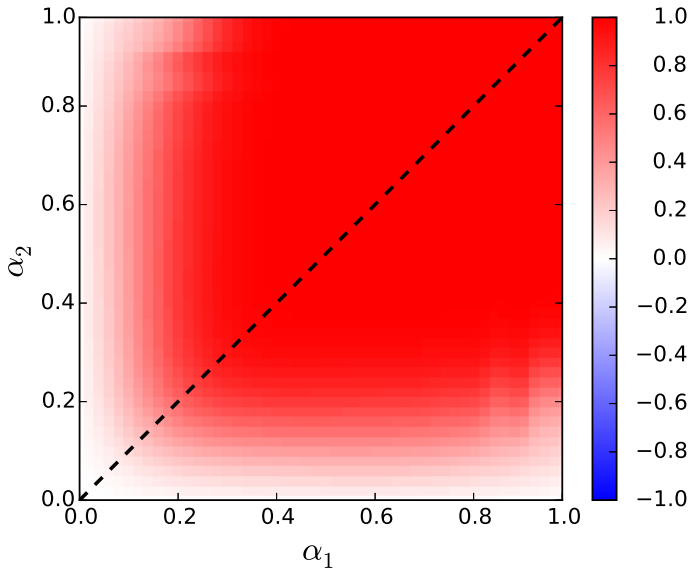



図 4.9: 有限マルコフ過程での検証（CC 優位解へ到達しやすい利得行列）

となる．ここから，次式をえる．

$$\begin{aligned}\Delta\gamma_1 &= \gamma_{1,\mathrm{C}} - \gamma_{1,\mathrm{D}} \\ &= P_{\mathrm{CC}}[f_1(\mathrm{CC}) - f_1(\varnothing)] + P_{\mathrm{CD}}[f_1(\mathrm{CD}) - f_1(\varnothing)] \\ &\quad + P_{\mathrm{DC}}[f_1(\varnothing) - f_1(\mathrm{DC})] + P_{\mathrm{DD}}[f_1(\varnothing) - f_1(\mathrm{DD})]\end{aligned}$$

$f_i(\varnothing) = 0$ を仮定した場合との対応関係は明らかである．このことから，$f_i(\varnothing)$ の値は報酬を平行移動させた程度の影響しかもたない．

### 4.8.1 CC 優位解の存在条件（無条件報酬あり）

CC 優位解の存在条件に関しても同じように置換すればよい．$f_i(\varnothing) \neq 0$ の場合，$\alpha_i \to 1$ において CC 優位解の存在条件は以下である．

$$f_1(\mathrm{CC}) > f_1(\varnothing) \quad \text{and} \quad f_2(\mathrm{CC}) > f_2(\varnothing)$$

## 4.9 議論

### 4.9.1 均衡解の存在条件とその個数

本章では，強化学習主体の囚人のジレンマを近似的に記述するモデルとしてベクトル場近似を定式化し，その近似モデルを用いて，存在する解の集合，解の存在条件，解の安定性などを解析した．その結果，相互裏切 DD に対応する均衡解（$P_i < 0.5$ をみたす安定点）は利得行列やパラメータ $\alpha_i$ に依らず常に存在することがわかった．また，パラメータ $\alpha_i \to 1$ の条件のもと，相互協調 CC に対応する均衡解（$P_i > 0.5$ をみたす安定点）が存在する必要条件は $f_i(\mathrm{CC}) > 0$ であるとわかった．また，本節では強化学習主体の囚人のジレンマでは少なくとも 1 つ，多くとも 4 つの均衡解が存在することを示した．それぞれの解は CC，CD，DC，DD に対応しているが，DD 優位解が常に存在するのに対し，CC 優位解は $f_i(\mathrm{CC}) > 0$，そして CD 優位解と DC 優位解は $f_i(\mathrm{DC}) > f_i(\mathrm{CD}) > 0$ で存在する（ただし $\alpha_i \to 1$）．これらの必要条件はゼロを基準とするが，ゼロとは選択しなかった行動に対して無条件で与えられる利得（無条件報酬）$f_i(\varnothing)$ のことだと示した．

### 4.9.2 協調行動の説明との関係

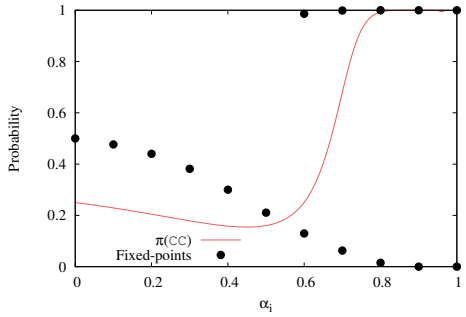
本来のモデルにおいて，この必要条件は「相互協調 CC の実現から与えられる利得が将来 C をだす確率を高める」と解釈できる．これは強化学習の枠組みにおいては自然な結果といえる．実際に，強化学習の枠組みでは，負の報酬 $f_i(\mathrm{CC}) < 0$ は行動 C をだす確率を低くし，同時に D をだす確率を高める．したがって，負の報酬 $f_i(\mathrm{CC}) < 0$ の場合，相互協調 CC の実現は相互裏切 DD を促進することになる．囚人のジレンマでは $f_i(\mathrm{CC}) > f_i(\mathrm{DD})$ ゆえに，逆のことも同時に起こる．すなわち，$f_i(\mathrm{CC}) < 0$ は $f_i(\mathrm{DD}) < 0$ を意味し，この条件では，相互裏切 DD の実現は相互協調 CC を促進することになる．以上から，負の報酬 $f_i(\mathrm{CC}) < 0$ の場合には強化学習戦略の行動はランダムに近いものとなる（図 4.6）．

本節で導出した CC 優位解へ到達しやすい利得行列（図 4.8）は図 4.6 の場合とは正反対の設定といえる．具体的には，相互協調 CC に対して正の報酬 $f_i(\mathrm{CC}) > 0$ を与えて C をだす確率を上げると同時に，相互裏切 DD に対して負の報酬 $f_i(\mathrm{DD}) < 0$ を与えて D をだす確率を下げる（すなわち C をだす確率を上げる）．このメカニズム的な記述からも，相

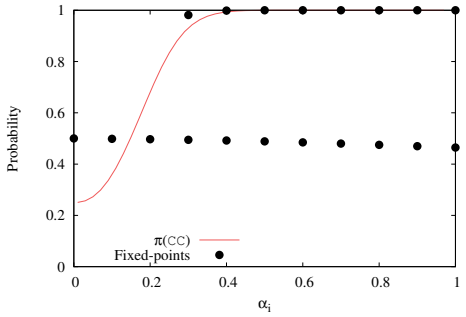
互協調の実現しやすさが伺える．このメカニズムは強返報性と呼ばれる協調行動の説明とも関係すると考えられる．強返報性とは，協調行動を盛大に賞める一方で裏切行動を盛大に罰するという，長期的な期待利益よりむしろ短期的な実益実害を重視することが，協調行動を持続させるには重要という考えである [21]．本章の結果は長期と短期のどちらか重要かという点には中立的であるが，強化学習を採用した場合，「協調行動を盛大に賞める一方で裏切行動を盛大に罰する」利得行列（図 4.8）ではより小さな $\alpha_i$ であっても，換言すればより短期的な経験に基づく場合でも，相互協調が可能であることを示している．事実，この利得行列を用いる場合，本来のモデルでも相互協調は容易に実現される．

### 4.9.3 有限マルコフ過程との相補性

ベクトル場近似は本来のモデル（エージェントシミュレーション）の局所的な挙動を記述することを目的としており，他方，その長期的な挙動すなわち定常分布を捉えることは必ずしもできない．実際，本来のモデルはベクトル場（近似モデル）上に存在するどれか 1 つの均衡に収束するとは限らず，ベクトル場を運動し続けている可能性もある．また，本来のモデルは履歴に依存した行動の選択を扱うが，ベクトル場近似は履歴を無視した平均的な行動の選択を扱い，本来のモデルが離散的な履歴ではとりえないような状態（累積利得）も含めて解となりうるかを分析している．図 4.10 は有限マルコフ過程の定常分布 $\pi(\mathrm{CC})$ とベクトル場近似の解 $P_i \in [0,1]$ を $\alpha_i$ の関数として図示した．これらの量は比較できないが，本来のモデルと近似モデルでの「相互協調 CC の実現されやすさ」を比較する意図で示した．有限マルコフ過程の分析によれば強化学習戦略は $\alpha_i \to 1$ で相互協調 CC をほぼ確率 1 で実現できるが，これはベクトル場近似の CC 優位解の出現と大まかに対応している．また，図 4.10 には示していないが，$f_i(\mathrm{CC}) < 0$ の場合では有限マルコフ過程でも相互協調 CC は偶然程度にしか実現されないことも確認できている．このように厳密な対応はないものの，相互協調の可能性を理解するという点では相補的な関係にあるといえるだろう．




利得行列 (3, 0, 5, 1)




4.7 節の利得行列

図 4.10: 有限マルコフ過程の $\pi$(CC) とベクトル場近似の解 $P_i$

### 4.9.4 先行研究の近似モデルとの関係

本章の結果は「$\alpha_i = 0.0$ ではランダムに振る舞うが，$\alpha_i \to 1.0$ にともない相互協調の実現確率が高まる」という前章の結果と整合的であると思われる．他方，この結果は「相互裏切すなわち $P_i < 0.5$ が唯一の安定点である」という別の近似モデルを用いた先行研究の報告 [31, 55] と矛盾する．Singh et al. [55] は利得行列の数値をそのまま勾配とするが，これは完全情報の信念学習に類似しており，その点では信念学習の結果とは整合的といえる．また，Kianercy and Galstyan [31] は Sato and Crutchfield [51] の連続時間結合レプリケータを用いて囚人のジレンマを解析した．ベクトル場近似の定式化は，導出過程は異なるが [51] と類似している[1]．その形式的な違いは，[51] によるモデルの“ 簡略化 ”[2] を行うか否かにある．この事実は，[51] における簡略化が近似モデルを本来のモデルとは定性的に異なるものに変質させた可能性を示している．

[1]Borgers and Sarin [8] や Sato and Crutchfield [51] の導出では相互学習のダイナミクスを記述する目的のもと，時定数の無限小化（連続化）を行っている．ところが，時定数の無限小化は「学習しない相手」と「一瞬」のうちに無限回の相互作用（ゲーム）が行われることを含意する．このことは，各プレイヤは上記の意味での「真の」利得行列を知れるばかりか，複数の主体が試行錯誤から同時に学習するという問題そのものが変質している．学習しない相手の戦略を既知とする学習モデルとも解釈できるが，これは信念学習と類似していると思われる．

[2]Sato and Crutchfield [51] は「簡単化のために期待報酬を行動確率の線形関数とする」と述べている．具体的には，$\Gamma_i$ に含まれる相互作用項 $P_1P_2$ を $P_1$ で置き換えている．

# 5　高次マルコフ戦略と今後の課題

## 5.1　本章の目的

強化学習戦略や１次マルコフ戦略など，過去の経験から意思決定を行う戦略クラスを統一的に記述する枠組みとして高次マルコフ戦略を定式化する．高次マルコフ戦略は利得関数の導関数を書き下すことができるため，進化などの個別の最適化技法に依存せずに，経験重視の意思決定を行う合理的プレイヤのゲームを扱い，その均衡解を分析できるという利点をもつ．本章ではまず強化学習戦略や進化ゲーム理論と，高次マルコフ戦略を扱うゲームとの関係を述べる．次に，高次マルコフ戦略に向けた最初の研究として１次マルコフ戦略を分析し，進化ゲーム理論でよく知られた戦略が Nash 均衡となる条件を調べる．高次マルコフ戦略は潜在的に高い記述力をもつと考えられるが，高次の場合を直接的に扱うには現状では技術的な問題が残されている．最後に，高次マルコフ戦略を用いた分析の今後の課題を論じる．

## 5.2　マルコフ戦略の無限回ゲーム

直前のゲームの結果（行動ペア）から次の行動を決める戦略を１次マルコフ戦略（以下，１次戦略）という．同様に，過去 $K$ 時点前までの行動ペアの履歴から次の行動を決める戦略を考えることができる．これを $K$ 次マルコフ戦略（以下，$K$ 次戦略）という．履歴に依存しない０次戦略は無条件戦略と呼ばれる．3 章で分析した強化学習戦略は $K$ 次戦略の一種であり，相手の行動履歴を利用せず，自分の行動履歴のみの関数として次の行動を確率的に決める[1]．3 章では，強化学習戦略の振る舞いが１次戦略のひとつ，しっぺ返し戦略 TFT と類似することを論じたが，このときの分析は複雑な高次戦略の振る舞いを直観的に捉えやすい１次戦

[1]本来，強化学習は無限次戦略である．近似の正しさは [26, 68] で論じている．

略の振る舞いで近似する．事実，強化学習はマルコフ決定過程と呼ばれるマルコフ過程の一種であり，強化学習と $K$ 次戦略との包含関係はマルコフ決定過程とマルコフ過程の包含関係と対応する．

$K$ 次戦略の集合を選択肢としてもつプレイヤのゲームを考えることができる．このゲームでは，一般的なゲームと同じく，各プレイヤは独立に戦略を選択し，両プレイヤの選んだ戦略の関数として各プレイヤへの利得が定まる．利得関数は選ばれた戦略ペアを無限回繰り返し対戦させたときの期待利得を返す（図 5.1）．このゲームは一般に無限回ゲームと呼ばれる．進化ゲーム理論における 1 次戦略ペアや本論文における強化学習戦略ペアは，この無限回繰り返しゲームのひとつの実現である（図 5.1）．進化ゲーム理論では生物個体を戦略で表し，各戦略の利得に依存して次世代の戦略の頻度分布が更新されるが，これは自然という 1 人のプレイヤ（セルフプレイ）によってなされる．他方，本論文 3 章ではこの意味での利得は重要ではなかったが，これは強化学習戦略を選ぶプレイヤが不在だからである．ただし，3.8 節では本章と同じく強化学習戦略 $\alpha_i \in [0, 1]$ を選ぶプレイヤを考えた．本章は明示的に 2 人のプレイヤが存在する無限回ゲームを扱う．

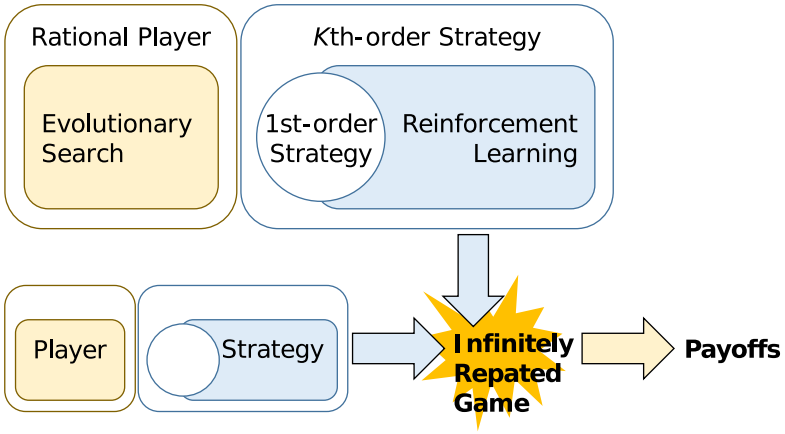



図 5.1: $K$ 次戦略の集合を選択肢とするゲーム，学習と進化の関係

1 次戦略のうち，直前の相手の行動のみに依存する戦略（応答戦略）は Nowak [41] により解析された．レプリケータを用いた進化ゲーム理論の分析は応答戦略 [45] から 1 次戦略 [46] に拡張され，より広範囲の戦略空間のなかで新しい進化的に安定な戦略が発見された．進化的最適化を用いず，しっぺ返し戦略 TFT にパラメータを導入し，導関数を最適化するアプローチは Molender [39] にみられる．Press and Dyson [47] は 1 次戦

略の利得関数を行列式で表現し分析している．本章ではこの技法を用い，強化学習戦略を含む一般的な $K$ 次戦略を記述する．ふたつの定式化の対応関係を示すため，まず次節では行列式を用いない定式化を復習し，次に後続の節では行列式を用いた定式化を述べる．

### 5.2.1 有限マルコフ過程と定常分布

プレイヤの数 $N = 2$，選択肢の数 $M = 2$ とする．$K$ 次マルコフ戦略を，過去 $K$ 時点前までの行動ペアの履歴 $X_i = ([X_i]_1, \ldots, [X_i]_K) \in \mathcal{M}^K$ のもとである行動 $a_l$ を選択する確率 $P_l(a_l|X_i)$ と記す．ある履歴 $X_j$ のもとで行動ペア $(a_1, a_2) \in \mathcal{M}$ が実現されたとき，次時点あるいは後続の履歴 $X_i$ の各要素 $[X_i]_k$ は $[X_i]_1 = (a_1, a_2)$ かつ $[X_i]_k = [X_j]_{k+1}$ $(1 < k < K)$ をみたす．その実現確率は $P(X_i|X_j) := P_1(a_1|X_j)\, P_2(a_2|X_j)$ となる．マルコフ戦略は状態空間 $X_i, X_j \in \mathcal{M}^K$ 上の遷移確率行列 $[Q]_{ij} = P(X_i|X_j)$ として表現でき[2]，初期確率分布 $\pi_0$ からの時間発展は $\pi_{t+1} = Q\,\pi_t$ となる．時間の極限 $t \to \infty$ で $\pi_\infty = Q\,\pi_\infty$ をみたし，$\pi := \pi_\infty$ を定常確率分布あるいは定常ベクトルという．ある履歴 $X_i$ の実現確率は $\pi_i$ で与えられる．ある履歴 $X_i$ に対応した利得を $r_i \in \mathbb{R}$ と記す．すべての履歴 $(X_1, \ldots, X_{M^{NK}})$ に対応した利得をまとめて $r = (r_1, \ldots, r_{M^{NK}})$ と記すとき，利得関数は $F(P_1, P_2, r) = \sum_i \pi_i r_i$ と書ける．

## 5.3 定式化

### 5.3.1 1次戦略

プレイヤ 1 の1次戦略を $p = (p_1, p_2, p_3, p_4)$，プレイヤ 2 の1次戦略を $q = (q_1, q_2, q_3, q_4)$ と記す．このとき，繰り返しゲームは遷移確率行列

$$Q^\top := \begin{pmatrix} p_1q_1 & p_1(1-q_1) & (1-p_1)q_1 & (1-p_1)(1-q_1) \\ p_2q_2 & p_2(1-q_2) & (1-p_2)q_2 & (1-p_2)(1-q_2) \\ p_3q_3 & p_3(1-q_3) & (1-p_3)q_3 & (1-p_3)(1-q_3) \\ p_4q_4 & p_4(1-q_4) & (1-p_4)q_4 & (1-p_4)(1-q_4) \end{pmatrix}$$

[2]ただし，$X_j$ から $X_i$ へ遷移不可能な場合は $[Q]_{ij} = 0$ とする．

で表せる．ここで，$Q^\top$ は行列 $Q$ の転置を表す[3]．この繰り返しゲームの定常状態は

$$\begin{aligned}&\pi = Q\,\pi\\ &\iff (Q-I)\,\pi = \mathbf{0}\end{aligned}\tag{5.1}$$

をみたす $Q$ の定常ベクトル $\pi$ で与えられる．

$Q^\top$ は，第 $i$ 列を $\textcircled{i}$，列の代入を $\leftarrow$ と記すとき，$\textcircled{2}' \leftarrow \textcircled{2}+\textcircled{1}$，$\textcircled{3}' \leftarrow \textcircled{3}+\textcircled{1}$，$\textcircled{4}' \leftarrow \textcircled{4}-\textcircled{1}+\textcircled{2}'+\textcircled{3}'$ の操作から，

$$\acute{Q}^\top := \begin{pmatrix} p_1q_1 & p_1 & q_1 & 1\\ p_2q_2 & p_2 & q_2 & 1\\ p_3q_3 & p_3 & q_3 & 1\\ p_4q_4 & p_4 & q_4 & 1\end{pmatrix}$$

をえる．ここで，行列式 $\det Q = \det Q^\top = \det \acute{Q}^\top$ をみたす．$(Q-I)^\top$ に同一の操作を適用して次式をえる．

$$\det(Q-I) = \begin{vmatrix} p_1q_1-1 & p_1-1 & q_1-1 & 0\\ p_2q_2 & p_2-1 & q_2 & 0\\ p_3q_3 & p_3 & q_3-1 & 0\\ p_4q_4 & p_4 & q_4 & 0\end{vmatrix}$$

$Q$ が非自明な $\pi$ をもつとき $\det(Q-I)=0$ をみたし，

$$(Q-I)\,\mathrm{adj}(Q-I) = \det(Q-I)I = \mathbf{0}\tag{5.2}$$

となる．ここで，$\mathrm{adj}(Q-I)$ は $Q-I$ の余因子行列である．(5.1)，(5.2) より，$Q$ の定常ベクトル $\pi$ と $\mathrm{adj}(Q-I)=(A_{ij})$ の各行は比例の関係

$$\pi \propto (A_{1k}, A_{2k}, A_{3k}, A_{4k})$$

にある．期待利得は $\pi=(\pi_1,\pi_2,\pi_3,\pi_4)$ と $r=(r_1,r_2,r_3,r_4)$ の内積

$$\begin{aligned}\pi\cdot r &= \pi_1r_1+\pi_2r_2+\pi_3r_3+\pi_4r_4\\ &\propto A_{1k}r_1+A_{2k}r_2+A_{3k}r_3+A_{4k}r_4\end{aligned}$$

[3] 3 章の遷移行列の表記と一致させるため，あえて転置している．

と書ける．$k = 4$ として次式をえる [47].

$$D(p,q,r) := \det \hat{Q}(p,q,r)$$

$$\hat{Q}(p,q,r) := \begin{pmatrix} p_1q_1 - 1 & p_1 - 1 & q_1 - 1 & r_1 \\ p_2q_2 & p_2 - 1 & q_2 & r_2 \\ p_3q_3 & p_3 & q_3 - 1 & r_3 \\ p_4q_4 & p_4 & q_4 & r_4 \end{pmatrix}$$

利得関数は次式で与えられる（分母は正規化項）[47].

$$F(p,q,r) := \frac{D(p,q,r)}{D(p,q,\mathbf{1})}$$

ここで，利得ベクトル $r$ はプレイヤ 1 とプレイヤ 2 で異なる．以降では，プレイヤ 1 の利得を $r$，プレイヤ 2 の利得を $u$ と書いて区別する．簡単のため，特別な理由のない限り $p$，$q$ を省略し，$F(r) := F(p,q,r)$，$D(r) := D(p,q,r)$，$\hat{Q}(r) := \hat{Q}(p,q,r)$ などと記す．

ゲームにおいて，プレイヤ 1 は $F(p,q,r)$ を戦略 $p$ に関して最大化し，他方，プレイヤ 2 は $F(p,q,u)$ を戦略 $q$ に関して最大化する．このとき，$F(p,q,r)$ と $F(p,q,u)$ を同時に最大化する $(p,q)$ を Nash 均衡という．

### 5.3.2 $K$ 次戦略

高次マルコフ戦略についても 1 次戦略と同様に定式化できる．本論文では 1 次戦略の分析のみを述べるため，ここでは省略する．

### 5.3.3 局所 Nash 均衡

そこで，本節では利得関数 $F$ の微分可能という性質を利用した分析を行う．利得関数 $F(p,q,r)$ が戦略 $p$ に対して微分ゼロという条件は Nash 均衡の必要条件を与える．すなわち，微分ゼロの条件をみたさないとき Nash 均衡ではないが，反対に，微分ゼロの条件をみたしても Nash 均衡とは限らない．導関数を用いた分析では戦略の局所的な変更を扱うため，$\epsilon$-近傍に対する Nash 均衡の拡張を定義する．局所 Nash 均衡は $\epsilon \to 1$ で Nash 均衡と一致する．

**Definition 5.3.1** (局所 Nash 均衡). 点 $p$ の $\epsilon$-近傍を $B(p,\epsilon) = \{p' \in [0,1]^4 : \|p-p'\| \leq \epsilon\}$ と記す．ここで，$\|\cdot\|$ はユークリッドノルムとする．ある $\epsilon > 0$ について, すべての $p' \in B(p,\epsilon)$ に対して $F(p',q,r) \leq F(p,q,r)$ をみたし，かつ，すべての $q' \in B(q,\epsilon)$ に対して $F(p,q',u) \leq F(p,q,u)$ をみたす戦略ペア $(p,q)$ を局所 Nash 均衡と定義する．

微分可能な利得関数 $F$ の場合，次の定義を与えられる．

**Definition 5.3.2** (微分可能な利得関数の局所 Nash 均衡). 微分可能な利得関数 $F(p,q,r)$ および $F(p,q,u)$ に対し，ある $\epsilon > 0$ について，すべての $p' \in B(p,\epsilon)$ に対して $\frac{\partial F(p,q,r)}{\partial p} \cdot (p'-p) \leq 0$ をみたし，かつ，すべての $q' \in B(q,\epsilon)$ に対して $\frac{\partial F(p,q,u)}{\partial q} \cdot (q'-q) \leq 0$ をみたす戦略ペア $(p,q)$ を局所 Nash 均衡と定義する．ここで，$\cdot$ はベクトルの内積を表す．

## 5.4 1次戦略の局所 Nash 均衡

1次戦略は $p,q \in [0,1]^4$ で定義されるが，本章では，進化ゲーム理論で扱われる代表的な1次戦略（表 3.1）を基礎として，1つの変数をもつ戦略のみを扱う．この場合, 各プレイヤのもつ1変数を除いて, $(p_1,p_2,p_3,p_4) = (q_1,q_3,q_2,q_4)$ をみたす．利得ベクトルに関しては $(r_1,r_2,r_3,r_4) = (u_1,u_3,u_2,u_4)$ をみたすものとする．囚人のジレンマでは $r = (f(\mathrm{CC}), f(\mathrm{CD}), f(\mathrm{DC}), f(\mathrm{DD}))$ である．本節では代表的な1次戦略（表 3.1）を2人ゲームの Nash 均衡という観点から分析する．

### 5.4.1 局所 Nash 均衡の数理解析

#### ALLC 対 ALLC

ALLC は $p = (1,1,1,1)$，$q = (1,1,1,1)$ と定義される．期待利得は $F(r) = f(\mathrm{CC})$ である．利得関数の1次微分は以下である．

$$\frac{\partial F(r)}{\partial p} = \big(f(\mathrm{CC}) - f(\mathrm{DC}), 0, 0, 0\big)$$

囚人のジレンマでは $f(\mathrm{CC}) - f(\mathrm{DC}) < 0$ である．プレイヤ 1 は，$p_1 + \frac{\partial F(r)}{\partial p_1} < 1$ より，$p_1$ を 1 から離れるように変化させる．ゆえに $p$ は局所 Nash 均衡ではない（$q$ も同様）．

### ALLD 対 ALLD

ALLD は $p = (0,0,0,0)$, $q = (0,0,0,0)$ と定義される．期待利得は $F(r) = f(\mathrm{DD})$ である．利得関数の 1 次微分は以下である．

$$\frac{\partial F(r)}{\partial p} = \big(0,0,0,f(\mathrm{CD}) - f(\mathrm{DD})\big)$$

囚人のジレンマでは $f(\mathrm{CD}) - f(\mathrm{DD}) < 0$ である．$p_1 \in [0,1]$ の制約の下では $\max(0,\ p_4 + \frac{\partial F(r)}{\partial p_4}) = 0$ ゆえにプレイヤ 1 は戦略を変化させない（$q$ も同様）．局所 Nash 均衡かどうかは近傍の点を調べる必要がある．

### RAND 対 RAND

RAND は $p = (1/2,1/2,1/2,1/2)$, $q = (1/2,1/2,1/2,1/2)$ で定義される．期待利得は $F(r) = (f(\mathrm{CC}) + f(\mathrm{CD}) + f(\mathrm{DC}) + f(\mathrm{DD}))/4$ である．利得関数の 1 次微分は以下である．

$$\frac{\partial F(r)}{\partial p} = \begin{pmatrix} [f(\mathrm{CC}) + f(\mathrm{CD}) - f(\mathrm{DC}) - f(\mathrm{DD})]/8 \\ [f(\mathrm{CC}) + f(\mathrm{CD}) - f(\mathrm{DC}) - f(\mathrm{DD})]/8 \\ [f(\mathrm{CC}) + f(\mathrm{CD}) - f(\mathrm{DC}) - f(\mathrm{DD})]/8 \\ [f(\mathrm{CC}) + f(\mathrm{CD}) - f(\mathrm{DC}) - f(\mathrm{DD})]/8 \end{pmatrix}$$

囚人のジレンマでは $f(\mathrm{CC}) + f(\mathrm{CD}) - f(\mathrm{DC}) - f(\mathrm{DD}) < 0$ である．ゆえに $p$ は局所 Nash 均衡ではない（$q$ も同様）．

### ALL$x$ 対 ALL$y$

ここまで変数を含まない戦略を調べたが，次に ALLC，ALLD，RAND を含む戦略クラスとして $p = (x,x,x,x)$, $q = (y,y,y,y)$ を考える．この分析では $x, y$ の 2 変数方向に関してのみ，局所 Nash 均衡の是否について情報をえられる．期待利得は以下である．

$$\begin{aligned} F(r) &= [(a-b)\,xy + bx + cy + d]/d \\ \text{where}\quad a &:= f(\mathrm{CC}) - f(\mathrm{DC}), \\ b &:= f(\mathrm{CD}) - f(\mathrm{DD}), \\ c &:= f(\mathrm{DC}) - f(\mathrm{DD}),\ \text{and} \\ d &:= f(\mathrm{DD}) \end{aligned}$$

利得関数の 1 次微分は以下である．

$$\frac{\partial F(r)}{\partial p} = \begin{pmatrix} x(ay^2 + by(1-y)) \\ x(ay(y-1) + b(y-1)^2) \\ (1-x)(ay^2 + by(1-y)) \\ (1-x)(ay(y-1) + b(y-1)^2) \end{pmatrix}$$

囚人のジレンマでは $a < 0$，$b < 0$ だからすべての軸で勾配は負であり，すべての $x > 0$ について $x \to 0$ すなわち ALLD へ向かう．ゆえに，$p$ は ALLD の場合を除いて局所 Nash 均衡ではないといえる（$q$ も同様）．

後続の節では変数 $x, y$ を導入した (G)TFT, WSLS, GRIM を分析する．

### TFT 対 TFT

TFT は直前の相手の行動を複製する戦略であり，$p = (1, 0, 1, 0)$ で定義される．TFT はノイズに弱いことが知られ，TFT 同士で対戦した場合，偶然に一度でも裏切 D をだすと際限ない相互裏切に陥る．これに対して，寛容な GTFT は相互裏切から抜けだす術をもち，具体的には相互裏切の場合でもある確率で協調 C をだす．そこで，この寛容性を変数として，一般化 TFT 戦略を $p = (1, x, 1, x)$，$q = (1, 1, y, y)$ と定義する．

一般化 TFT 戦略では，期待利得は $F(r) = F(u) = f(\mathrm{CC})$ である．この結果は，$x, y$ に依存せず相互協調となることを意味する．他方，利得関数の 1 次微分は $x, y$ に依存し，以下である．

$$\frac{\partial F(r)}{\partial p} = \left( \frac{(a-b)y - (2a-b-c)}{xy - x - y}, 0, 0, 0 \right)$$
$$\text{where} \quad (a, b, c, d) := (f(\mathrm{CC}), f(\mathrm{CD}), f(\mathrm{DC}), f(\mathrm{DD}))$$

ここで，$0 \leq x, y \leq 1$ より分母は $xy - x - y \leq 0$ である．したがって，傾きの符号は相手のパラメータ $y$ と利得 $r$ に依存する．分子を $\phi(y)$ と記す．囚人のジレンマの条件から $\phi(0) < 0$，$\phi(1) > 0$ である．$\frac{\partial F(r)}{\partial p_1} < 0$ ならば，局所 Nash 均衡ではない．その条件は

$$\begin{aligned} &\phi(y) = (a-b)y - (2a-b-c) > 0 \\ &\implies c > (2a-b) - (a-b)y \end{aligned}$$

である．これを $a$ の関数 $c = \xi_y(a) = (2a-b) - (a-b)y$ とみなす．囚人のジレンマの条件と併せて図 5.2 をえる．囚人のジレンマの条件 $c > a > b$

かつ $2a > b+c$ から $c$ のとりうる範囲は $a < c < 2a-b$ であるが，これは $a = \xi_1(a) < \xi_y(a) < \xi_0(a) = 2a-b$ と一致する．$a \le c \le \xi_y(a)$ の範囲で $p$ は局所 Nash 均衡ではない（ALLC を含む）．他方，$\xi_y(a) < c < 2a-b$ の範囲で $p$ は傾き 0 となる（TFT を含む）．境界 $\phi(y) = 0$ をみたす $y$ において，$r = (3, 0, 5, 1)$ のとき $y = 1/3$ すなわち GTFT をえる [39]．

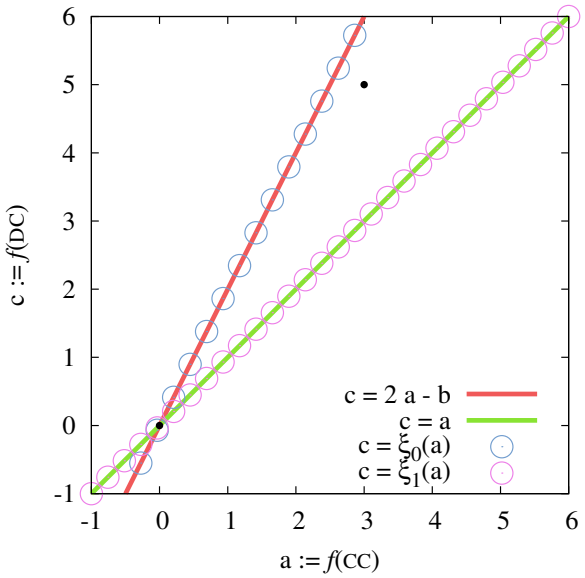



図 5.2: $f(\mathrm{CC})$ と $f(\mathrm{DC})$ の関係（TFT 対 TFT）．$(a, b, c, d) := (f(\mathrm{CC}), f(\mathrm{CD}), f(\mathrm{DC}), f(\mathrm{DD}))$．囚人のジレンマは $a < c < 2a - b$ をみたし，赤線と緑線の間の領域のみ．一般化 TFT は $c > \xi_y(a)$ の範囲で局所 Nash 均衡ではない．$y$ は相手のパラメータ

### WSLS 対 WSLS

WSLS は痛み（DD，CD）に対して行動を切り換え，喜び（CC，DC）に対して前回と同じ行動をとる戦略であり，$p = (1, 0, 0, 1)$ と定義される．そこで，痛みに対して行動を切り換える確率を変数として，一般化 WSLS 戦略を $p = (1, x, 0, 1-x)$，$q = (1, 0, y, 1-y)$ と定義する．期待利得は $F(r) = F(u) = f(\mathrm{CC})$ である．利得関数の 1 次微分は以下である．

$$\frac{\partial F(r)}{\partial p} = \left( \frac{(a-b)y + (2a-c-d)}{xy - x - y + 1}, 0, 0, 0 \right)$$

where $\quad (a, b, c, d) := (f(\mathrm{CC}), f(\mathrm{CD}), f(\mathrm{DC}), f(\mathrm{DD}))$

ここで，$0 \leq xy - x - y + 1 \leq 1$ より，局所 Nash 均衡でない条件は

$$\begin{aligned} &\phi(y) = (a-b)y + (2a - c - d) < 0 \\ &\Longrightarrow\ c < (2a - d) + (a - b)y \end{aligned}$$

これを $c = \xi_y(a) = (2a - d) + (a - b)y$ とみなす．囚人のジレンマの条件と併せて図 5.3 をえる．$\xi_0(a) < c < 2a - b$ の範囲では一般化 WSLS は局所 Nash 均衡ではない．$\xi_0(a) < c$ とは $2f(\mathrm{CC}) < f(\mathrm{DC}) + f(\mathrm{DD})$ であり，協調を繰り返すよりも搾取＋裏切のほうが高い利得を与える場合である．したがって，一般化 WSLS の局所 Nash 均衡の是否は利得のみから定まる場合があり，$2f(\mathrm{CC}) < f(\mathrm{DC}) + f(\mathrm{DD})$ の場合は相手の $y$ によらず局所 Nash 均衡ではない．

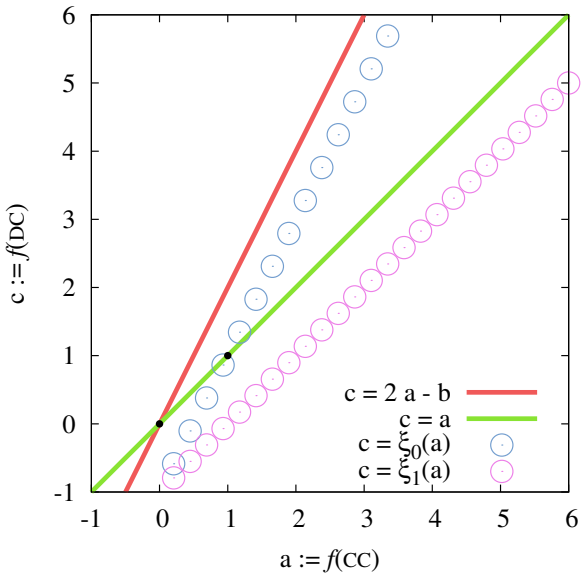



図 5.3: $f(\mathrm{CC})$ と $f(\mathrm{DC})$ の関係（WSLS 対 WSLS）．$(a, b, c, d) := (f(\mathrm{CC}), f(\mathrm{CD}), f(\mathrm{DC}), f(\mathrm{DD}))$．囚人のジレンマは $a < c < 2a - b$ をみたし，赤線と緑線の間の領域のみ．一般化 WSLS は $c > \xi_y(a)$ の範囲で局所 Nash 均衡ではない．$y$ は相手のパラメータ

### GRIM 対 GRIM

GRIM は相手が一度でも裏切 D を選べば以降常に D をだすという戦略であり，1 次戦略としては $p = (1, 0, 0, 0)$ と定義される．そこで，裏切の連

鎖から抜けだす寛容さを変数として, 一般化 GRIM 戦略を $p = (1, x, x, x)$, $q = (1, y, y, y)$ と定義する．期待利得は $F(r) = F(u) = f(\mathrm{CC})$ である．利得関数の 1 次微分は以下である．

$$\frac{\partial F(r)}{\partial p} = \left( \frac{-(b-d)(1-y)x - (c-d)y + (a-d)}{xy}, 0, 0, 0 \right)$$
$$\text{where} \quad (a, b, c, d) := (f(\mathrm{CC}), f(\mathrm{CD}), f(\mathrm{DC}), f(\mathrm{DD}))$$

ここで，$0 \leq xy \leq 1$ より，局所 Nash 均衡でない条件は

$$\phi(x, y) = -(b-d)(1-y)x - (c-d)y + (a-d) < 0$$
$$\Longrightarrow \; c < [-(b-d)(1-y)x + dy + (a-d)]/y$$

これを $\xi_{xy}(a) = [a - (1-y)(bx + d(1-x))]/y$ とみなす．囚人のジレンマの条件と併せて図 5.4 をえる．$\xi_{xy}(a)$ の切片は $x$ に依存するが，傾きは $x$ に依存せず，$\xi_{x0}(a) \approx bx + d(1-x)$ （垂線），$\xi_{x1}(a) = a$ である．したがって，$y \to 1$（ALLC）ではどのような $f(\mathrm{CC})$，$f(\mathrm{DC})$ でも $p$ は局所 Nash 均衡ではないが，他方，$y \to 0$（GRIM）ではどのような $f(\mathrm{CC})$, $f(\mathrm{DC})$ でも $p$ は傾き 0 となる．

### 5.4.2 局所 Nash 均衡の数値検証

本節では，進化ゲーム理論でよく知られた 1 次戦略が局所 Nash 均衡となりうるかを 1 次微分を使って分析してきた．その結果，ALLD，TFT，GRIM は局所 Nash 均衡となる可能性をもち，他方 ALLD，RAND は常に局所 Nash 均衡とはなりえず，また WSLS は $f(\mathrm{CC})$，$f(\mathrm{DC})$ によっては局所 Nash 均衡となりえないことがわかった．ALLD，TFT，WSLS，GRIM が局所 Nash 均衡となるかは定義 5.3.1 か定義 5.3.2 を用いた分析を要する．本節で解析的に求めた期待利得を用い，これらの定義に基づく分析を数値的に行った．数値実験ではさまざまな $\epsilon$ を用いたが，局所 Nash 均衡に関して同じ判定結果をえた．その結果，ALLD，TFT，GRIM は異なる $f(\mathrm{CC})$，$f(\mathrm{DC})$ に対しても Nash 均衡であることが示唆された．つまり，両プレイヤがこれらの戦略をとるとき，片方のプレイヤは独立に戦略を変えることで期待利得を高めることができない．他方，WSLS は先述の予想どおり，$2f(\mathrm{CC}) < f(\mathrm{DC}) + f(\mathrm{DD})$ をみたす利得行列では，独立に戦略を変えることで期待利得を高めることが可能であり，Nash 均

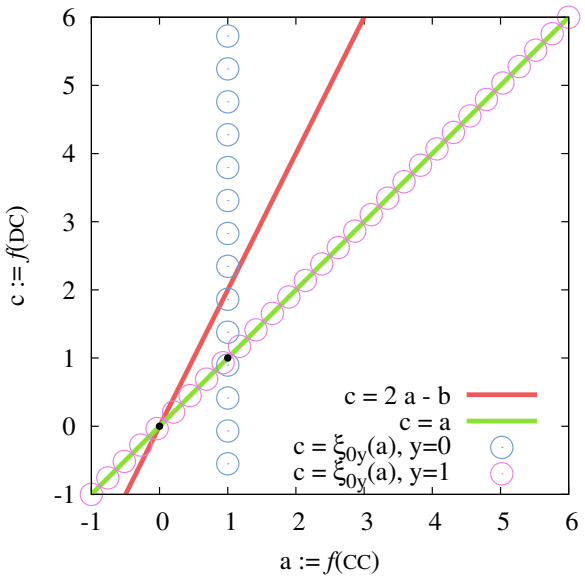



図 5.4: $f(\mathrm{CC})$ と $f(\mathrm{DC})$ の関係（GRIM 対 GRIM）．$(a, b, c, d) := (f(\mathrm{CC}), f(\mathrm{CD}), f(\mathrm{DC}), f(\mathrm{DD}))$．囚人のジレンマは $a < c < 2a - b$ をみたし，赤線と緑線の間の領域のみ．一般化 GRIM は $c > \xi_{xy}(a)$ の範囲で局所 Nash 均衡ではない．$x$ は自分，$y$ は相手のパラメータ

衡ではないことが示唆された．その他の場合は Nash 均衡であることが示唆された．

ALLD，TFT，WSLS，GRIM などは $p \in \{0, 1\}^4$ からなる決定論的戦略であるが，$p \in (0, 1)^4$ からなる確率論的戦略についても数値的に調べた．このさい，定義 5.3.1 か定義 5.3.2 に基づく分析のほかに，攪乱をともなう勾配降下法による最適化も行った．その結果，十分に小さな $\epsilon$ においても局所 Nash 均衡と判定されたのは ALLD と GRIM に近い $p \in (0, 1)^4$ をみたす戦略のみで，TFT と WSLS に近い戦略は TFT，WSLS を表す点から離れ ALLD へと向かう．ALLD，GRIM に近い確率論的戦略の期待利得はほぼ $f(\mathrm{DD})$ で，相互裏切に陥ることが示唆される．

## 5.5 議論

### 5.5.1 １次戦略とその背景にある理論の関係

GRIM，TFT，WSLS はそれぞれ異なる理論を１次戦略として表現したものである．GRIM は素朴定理を原点とし，TFT はメタゲーム理論を原点とするが，いずれの理論も利得行列の数値によらずに相互協調の可能性を示している．本章の分析からも，これらの戦略は利得行列の数値に依存せず，しかし相手の戦略には依存して，相互協調が可能であることが示唆された．他方，WSLS は要求学習を原点とするが，その振る舞いは利得行列の数値から影響を受ける．本章の分析からは局所 Nash 均衡でなくなる条件として $2f(\mathrm{CC}) < f(\mathrm{DC}) + f(\mathrm{DD})$ という関係式が導かれたが，この関係式は経験に基づき行動を決める学習戦略にとっては裏切の誘因を特徴づけている．強化学習戦略もパラメータによってはこの関係式の影響を受けることが予想される．また，確率論的戦略の場合，WSLS に近い戦略から ALLD に近い戦略へ向かうことが示唆されたが，同じく確率的な強化学習戦略は ALLD に対してはわずかに ALLD よりも高い期待利得をえられない (3 章)．つまり，強化学習戦略と ALLD 戦略を選択肢とするゲームを考えた場合，わずかの差で ALLD を選択する誘因が存在する．

このように，圧縮された表現である１次戦略がその背景にある理論の性質を少なからず継承していることは，広大な戦略空間をもつゲームの全体像を理解するうえで重要な意味をもつと思われる．素朴定理やメタゲーム理論など本来経験に基づく意思決定方式でないものを扱うには限界があると思われるが，過去の経験に基づき確率的に意思決定を行うマルコフ戦略に関しては１次戦略による近似的理解の有用性を示唆する．

### 5.5.2 高次戦略を理解するための１次戦略

高次マルコフ戦略は戦略の記述が $K$ の指数で増加するため，$K$ が大きい場合ほどその定常状態における振る舞いを理解することが難しくなる．そこで，本論文 3 章では高次マルコフ戦略の一種である強化学習戦略の振る舞い（１次の定常分布）を１次戦略で近似的に理解することを試みた．本論文 3 章では，強化学習戦略の振る舞いを高精度で再現する複数の１次戦略が見つかり，その平均的な戦略を求めることで，強化学習戦略が相互協調を実現するとき TFT 的な戦略であることや，パラメー

タに依存して RAND，WSLS，TFT など異なる 1 次戦略的振る舞いを習得しているという知見をえられた．しかしながら，この理解の方法は発見的であり，有用なアイデアではあるものの，高次マルコフ戦略である強化学習が長期的な意味で TFT 的である可能性を示唆するにすぎない．

より厳密に高次戦略のゲームを理解するための鍵は $K$ 次戦略のゲームと $K+1$ 次戦略のゲームがどのように関係しているかを理解することにあると思われる．その第 1 段階は，0 次戦略と 1 次戦略の関係であると考えられる．0 次戦略すなわち標準ゲームの混合戦略では唯一の Nash 均衡は ALLD ということが知られている．数値実験の結果から，確率論的な $p \in (0,1)^4$ なるすべての戦略は，局所的にみれば近傍により高い期待利得を与える戦略が存在し，大域的にみれば ALLD か GRIM へ向かって切り換えられる誘因をもつことが示唆される．また，確率論的な GRIM は振る舞いとしては ALLD となっている．以上から，確率論的な 1 次戦略の全体を扱う場合，0 次戦略と同じく，唯一の Nash 均衡は ALLD であるという仮説が考えられる．この仮説を理論的に示すことは今後の課題のひとつである．

他方で，本章の分析から，決定論的な 1 次戦略 $p \in \{0,1\}^4$ が Nash 均衡となる可能性が示唆された．また，理論的な解析から，完全に決定論的でなくとも，一部に $p_i \in \{0,1\}$ をもつ戦略が 1 次戦略全体を扱う IPD の Nash 均衡となる可能性がある．マルコフ戦略の相互協調に関しては，とくに (a) $p_1 = q_1 = 1$ すなわち相互協調を長期的に持続できる，(b) $p_4 \approx q_4 \approx 0$ すなわち相互裏切に対して寛容すぎない，といった条件の重要性が示唆される．条件 (a), (b) を両方みたす TFT，GRIM は利得行列の数値によらず, Nash 均衡となることが示唆された．他方, (a) をみたし，(b) をみたさない WSLS は裏切の誘因の高い $2f(\mathrm{CC}) < f(\mathrm{DC}) + f(\mathrm{DD})$ をみたす利得行列において Nash 均衡ではなくなる．

現状，高次マルコフ戦略の理解は定性的なものとなるだろう．そのため，本章では高次マルコフ戦略の構造を示しながらも，第 1 段階として 1 次戦略の定性的・定量的解析を示した．高次マルコフ戦略のゲームを捉えるには現時点では道具立ても不十分であるが，段階的な発展の先には，強化学習を含む高次マルコフ戦略全体を扱うゲームにおいて，強化学習戦略の位置づけや，相互協調すなわちジレンマ解消の条件をより詳細に明らかにできると期待する．

# 6　総合考察

本章では，協調問題の解決という観点から，本論文の主要な結果を先行研究の知見と対比しながら整理する．そのうえで，今後の発展として，協調の計算論的理解に向けて本論文の主要な知見をまとめる．最後に，実社会への応用可能性について論じる．

## 6.1　本論文の主題とモデル

本学位論文では，情報の不確実性をともなう繰り返し囚人のジレンマを扱い，過去のゲームの経験から行動を調整する強化学習プレイヤの間で協調問題が解決される条件を探ってきた．本論文で取り入れた不確実性や学習といった要因は心理学実験に着想をえている．囚人のジレンマを扱った心理学実験では，被験者は繰り返しゲームをプレイするなかで学習することや，ゲームに関する情報が欠如しているときに協調しやすいなどの事例が報告されている．

本論文で扱った不確実性とは，各プレイヤの視点からはアクションに対するフィードバックが確率的であり，その確率分布に関する情報を各プレイヤが明にもたないことを意味した．この不確実性は問題状況（すなわち囚人のジレンマ）に関して各プレイヤが部分的な情報しかえられず，各プレイヤの視点からは問題状況の全体像を把握できないという設定に由来する．より具体的には，本論文では，理論的な観点から極端に情報の限定された状況として，問題状況を記述する利得行列や，（ともに）問題に取り組む相手の行動を不可視とし，各プレイヤが自分の行動（アクション）とそれに対する利得（フィードバック）しか情報としてもたない状況を扱った．

本論文で扱った強化学習とは，上記のような極端に情報の限定された状況にも適用できる学習アルゴリズムのひとつである．計算目的としては，強化学習は一般に利益の最大化を目的関数とし，利益を高めるように行動確率を調整する．ここで，利益とは（長期的な）期待利得を意味

する．上記のように情報の限定された状況では，強化学習プレイヤは自己に関する情報しかもたず，ゆえに他者の利益などに関わらず，自己利益を最大化するように学習するほかない．アルゴリズムとしては，強化学習は各行動に対する累積利得に比例した確率で行動を選択し，その行動に対してえられた利得を用いて累積利得を更新する，という処理を繰り返す．このとき，累積利得の計算には割引率（記憶保持率）パラメータ $\alpha_i$ を用い，過去の利得ほど大きく割引される（$\alpha_i = 1$ で割引なし）．累積利得とは，強化学習にとっては過去のゲームの経験（記憶）を情報圧縮した表現である．そのため，このパラメータ $\alpha_i$ は強化学習プレイヤの考慮できる実効的な履歴の長さと関連し，プレイヤが保持可能な累積利得の上限いわば記憶容量の上限を定める．

学習戦略をとる 2 人のプレイヤのゲームでは学習対象（あるいは問題）も動的に変化するため，一般に，同時学習の帰結（行動確率の定常分布）は自明ではないと考えられる．とくに囚人のジレンマのように，標準型ゲームにおいて集団最適（Pareto 効率）と個人最適（Nash 均衡）が一致しないゲームでは，その展開型ゲームにおける行動確率の定常分布は Nash 均衡の予想と一致しない可能性がある [1]．本論文で扱った不確実な囚人のジレンマは，ゲーム（社会的問題状況）を通して間接的に繰り返し相互作用する 2 人の学習プレイヤの意思決定をモデル化しているものと捉えられる．本論文では，十分に長期的な履歴から学習するとき，強化学習プレイヤは相互協調（Pareto 効率）を高い確率で実現できる場合があることを示した．

## 6.2 協調問題の解決可能性

### 6.2.1 情報の次元と時間の次元

図 6.1 は，既存の相互協調の可能性（2 章）を「情報の次元」と「時間の次元」から張られる平面上に整理したものである．時間の次元は長期的過去から短期的過去，短期的将来から長期的将来を配置する軸である．他方，情報の次元は完全情報・完備情報から不完全・不完備情報までを配置する軸である．素朴定理は完全・完備情報のもと長期的将来の期待利益を計算して行動を決めるため，図の右上に位置する．有限回繰

[1] 対照的に，例えば，じゃんけんゲームではグー，チョキ，パーの行動確率は長期間平均をみれば 1/3 になるという予想が立つだろう．

り返しゲームを扱う後方帰納は素朴定理よりも時間の次元上で中央寄りに位置づけできる．信念学習は完全・完備情報のもと長期的過去の合計利益を計算して行動を決めるため，図の左上に位置する．強化学習は不完全・不完備情報のもと長期的過去の期待利益を計算して行動を決めるため，図の左下に位置する．素朴定理，信念学習，強化学習のもつ「割引率パラメータ」は時間の次元上での幅と対応づけられる．他方，情報の次元に関しては，各推論方式，学習方式の定義から定まる部分が大きくあまり幅をもたない．最後に，進化ゲーム理論で典型的に用いられる１次戦略は，その理論的背景を考慮した場合，平面の中央付近に位置づけできると思われる．１次戦略は過去の結果から行動を決めるが，TFT や GRIM のように，１次戦略に単純化される前の理論は将来の予想から行動を決めるものを含むため，中央付近に幅をもたせた．

図 6.1 には各理論（推論，学習）がセルフプレイ時に相互協調 CC，相互裏切 DD を実現しえるかどうかを合わせて示した．素朴定理は相互協調や相互裏切を含むすべてを実現しうるため，相互協調の可能性に対して示唆的であるものの，協調問題を直接的に解決しているとはいえない．進化ゲーム理論ではセルフプレイ時に相互協調を実現できる１次戦略が知られているが，すべての他の１次戦略の侵略を防ぐことはできず，またノイズのある確率的な環境で相互協調を維持することは難しい．信念学習は経験からの学習という観点を採用しているが，利得行列を参照して相手の行動に対する最適応答を学習するため，学習の結果として相互裏切のみが実現される．これらに対して，本論文で採用した強化学習は利得行列の不可視という限定合理性のもと，互いに十分に長期的な経験から学習する場合，学習の結果として相互協調が実現される．

図 6.1 から，協調問題に関する他の理論の体系での強化学習戦略の位置づけ，および，強化学習戦略が相互協調を実現できたポイントを考察する．第１に，強化学習と素朴定理は，経験重視（過去）か予想重視（将来）かで異なるが，どちらもより長期的利益を考慮した場合ほど相互協調を実現しやすいという共通点をもつ．第２に，強化学習と信念学習は，どちらも経験重視という点で一致しているが，利得行列に関する不確実性で異なり，利得行列を不可視とする強化学習のみが相互協調を実現できる．第３に，強化学習と１次戦略は，どちらも過去の行動履歴から次の行動確率を決めるが，その戦略を記述する履歴の長さで異なり，より長い履歴に基づく強化学習戦略は１次戦略に対して最適に振る舞うことができる．以上の対比から，利得行列（利害関係）に関して不確実な状

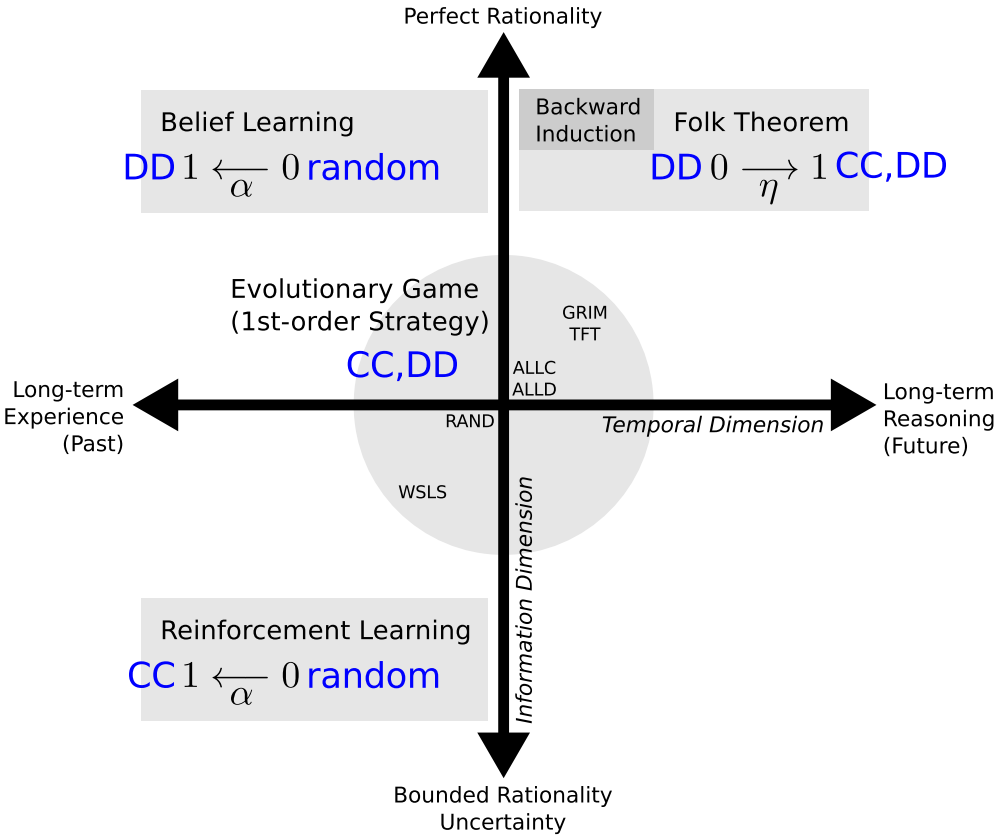



図 6.1: 相互協調の可能性の情報の次元と時間の次元による分類

況で，長期的な過去の経験から，長期的な自己利益を最大化するという計算目標が，より一般的に，強化学習戦略が相互協調を実現できた理由だと考えられる．

### 6.2.2 副産物としての返報性

協調問題を扱った研究では返報性（しっぺ返し戦略）自体を協調のメカニズムとすることが多い．返報的行動は自然選択により主体の意図とは無関係な行動パタンとして形成されるか（進化ゲーム理論），認知主体の自他入れ子型推論の帰結として発現する（メタゲーム理論）と考えられてきた．メタゲーム理論 [27] の考え方では，各認知主体の目的関数は仮説的な相手の目的関数を内包したものとなっており，相互協調が約束されるには自他ともにこの推論を行うことが前提となる．Axelrod [3] はしっぺ返し戦略の性質を明示的に相手と関連づけて論じている．また，協調

問題の解決には他者の情報が必要という考えは強化学習をもちいた研究 [6, 37] からも支持されている．これに反して，本論文の結果は，他者の情報は必要ではなく，各認知主体が自己利益という目的関数を最大化した結果として，確率的な状況でも相互協調を実現できる返報的行動が発現することを示している．この知見は，直接的に返報的行動を導く自他入れ子型推論的な意味づけに対して，自然選択の結果という意味づけに類した，最適化の副産物としての返報的行動といえる．返報的行動が利己的行動の副産物として生じえるならば，本論文の知見は協調行動の進化における疑問 [3, 24, 25] を学習レベルで解消できるという示唆を協調の進化に関して与える．

### 6.2.3 協調促進メカニズム

ゲーム的状況にある主体に強化学習を想定できる場合に，部外者・管理者の立場から，どのような利害構造を与えれば相互協調を促進できるかを論じる．本論文の結果から，選択していない行動に対して無条件で与えられる報酬を $f(\varnothing) = 0$ とするとき，相互協調の報酬 $f(\mathrm{CC}) > f(\varnothing) = 0$ がまず必要条件である．これは望ましい行動を促進するよう正の報酬を与えることを意味する．強化学習主体の意思決定には累積利得の差分が本質的なので，相互裏切の報酬 $f(\mathrm{DD})$ との差異 $f(\mathrm{CC}) - f(\mathrm{DD})$ が効いてくる．実際，$f(\mathrm{DD}) < f(\varnothing) = 0$ の場合，裏切 D を選択することは協調 C の選択を促す．これは望ましくない行動を抑制する負の報酬を与え，その結果として，望ましい行動を促進することを意味する．したがって，$f(\mathrm{CC}) > 0$ かつ $f(\mathrm{DD}) < 0$ をみたすような利害構造を指定できる仕組みが存在すれば，それは協調促進メカニズムといえる．この条件をみたすとき，短期的な利益を指向する場合でも相互協調が実現されやすくなる．このメカニズムは強返報性 [21] と呼ばれる協調行動の説明とも一致すると考えられる．

### 6.2.4 協調の計算論に向けて

本節では，今後の発展として，協調の計算論的理解に向けて本論文の主要な知見をまとめる．計算論のレベルの記述は認知主体の計算目標を記述する（最適化問題では目的関数がこれに含まれる）．先述のとおり，本論文で扱った不確実な囚人のジレンマは，ゲーム（社会的問題状況）を

通して間接的に繰り返し相互作用する２人の学習プレイヤの意思決定をモデル化しているものと捉えられる．強化学習主体は，標準的なゲーム理論の想定と同じく，個人的な利得の期待値すなわち自己利益の最大化を目的とする．本論文では，この計算目標をもつ強化学習は，十分に長期的な履歴を考慮した場合，相互協調を高い確率で実現できることを示した．しかし，ある認知主体が静的な環境と相互作用する際の情報処理あるいは「計算」を記述する場合とは異なり，本論文で扱った問題のように複数の認知主体が相互作用し同時に学習する場合には，個別主体の計算目標を記述するだけでは協調問題に対する計算論のレベルの記述としては不十分だろう．言い換えれば，相互協調に対する計算論のレベルの記述は２人のプレイヤがどのような計算目標を「同時に」達成しているかを記述する必要があると思われる．強化学習の観点からのアプローチとしては，強化学習を $K$ 次戦略と捉え，$K$ 次戦略と $K+1$ 次戦略との関係を分析する（5 章）など，新しい視点からの分析が必要となるだろう．他方で，本論文でえられた知見から示唆をえて，強化学習に依らないより単純な数理モデルを導出するアプローチもありえる．

本論文でえられた知見から，相互協調の計算論のレベルに向けて，以下の示唆がえられた．まず，自己利益の最大化という「個人的な計算目標」のもとで相互協調が実現されうる．このことは，個人の計算目標を記述することは，複数の「計算」の相互作用を記述するうえで出発点となるだろう．次に，自己に関する情報のみという極端に情報の限定された状況においても相互協調が実現されうる．このことは，個別主体の意思決定モデルが非常に単純なもので済むことを示唆する．また上記を併せて，本論文の扱う問題状況がそうであるように，協調問題の解決に関するひとつの描像は「互いを知らない ２人の学習主体それぞれが，囚人のジレンマという社会的問題（ゲーム）をそれと知らず，自分と動的な環境との問題として解決しようとした帰結」というものである[2]．この描像は，本論文よりも単純化されたモデルとして，各プレイヤが自分に割り当てられた２本腕スロットマシンに対して行動を最適化する（ただし，すべてのスロットマシンは相互依存関係にある）という古典的な強化学習問題のゲーム理論的な拡張として定式化できる可能性を示している．この描像のもとで，相互協調に至る道筋を計算目標として示すことができれば，協調を計算論のレベルで記述できたといえるだろう．

[2] 1.3 節で論じたが，本論文で扱った囚人のジレンマは複数の主体が関与し，そのため個別主体には問題の全体像を把握できなような複雑な問題状況を，情報の限定性という性質を残したまま単純化したものと解釈できる．

## 6.3 実社会への応用可能性

協調問題は，タダ乗り問題などの根底にある問題であり，日常経験する社会的な状況，企業における組織的活動や社会的実践にともなう．タダ乗り問題は「他人がやるなら自分はやらない」という考えが「誰もやらない」という帰結をもたらす状況を捉える．ゲーム理論において，タダ乗り問題は囚人のジレンマの多人数版である共有地の悲劇や公共財ゲームとして定式化されている．どちらのゲームも，本質的には個人からの投資と集団への収益分配との関係を扱い，ある一定人数（一定金額）以上の投資があれば事業が成功し集団として正の収益をえられるが，多数の他人が投資するならば自分は投資せずとも利益だけをえられるため，タダ乗りの誘惑が存在する．本学位論文では，限定的ではあるが，強化学習戦略を用いた共有地の悲劇や公共財ゲームの結果を付録 A に含めた．これらの分析結果は，本学位論文でえられた知見と整合的であり，強化学習戦略が囚人のジレンマの拡張である公共財ゲームにおいても，個人最適（Nash 均衡解）ではなく，集団最適（Pareto 効率解）へ到達可能な場合があることを示している（詳細な条件は今後の課題とする）．換言すれば，学習可能な状況においては，強化学習戦略は個人最適と集団最適の対立という意味での協調問題を解決できる．

Cabrera and Cabrera [15] は企業・組織でみられる協調問題をまとめている．組織学・心理学の知見によれば，協調問題の未解決すなわち裏切行為の黙認は，裏切行為を次第に普及させ，集団全体の活力を低下させるなど，組織的活動を停滞させる原因となる．経営学の知見によれば，企業活動では問題点や解決策の共有が作業能率の促進という利益を生むが，情報を共有する作業には労力（コスト）がともなうゆえにある種の協調問題（タダ乗り問題）が生じる．Cabrera and Cabrera [15] による上記の知見から，協調問題の解決は組織的活動を円滑に実現するための必要条件であり，協調問題の解決法を理解することは組織的知識創造活動を促進する方法を理解することに繋がると考えられる．一般にこれらの状況では問題状況の把握や情報共有の不足などが問題とされるが，その際，情報の種類に注意を払うことが重要であると思われる．古典ゲーム理論の知見によれば，意思決定主体が利害関係について十分な情報をもつ場合ほど協調問題は解決できない．一方で，本学位論文の知見は利害関係に関する不確実性が高い場合には協調問題が解決されうることを示している．換言すれば，個別主体が部分的な問題に対する自らのアクションに応じたフィードバックを手がかりに行動すれば，利害対立の全体像を

把握していない場合にむしろ利害対立を克服し，組織的協調あるいは集団利益を最大化する状態の実現が可能であることを示唆する．

# 7　結論

## 7.1　まとめ

本学位論文では，協調問題すなわち「各主体が個人として最も望ましい状態を追求すると，個人の行為の寄せ集めとして，集団として最も望ましい状態が実現されない」という個人最適と集団最適の対立を問題とした．先行研究の知見から，「不確実な状況における学習」が協調問題を解決する要因であると考え，協調問題を「利害関係や他者の情報に関する不確実性の高い問題状況」と捉え，不確実性に対処する認知的な能力として「学習」，具体的には強化学習を採用した．この観点から，本学位論文では，強化学習を行う主体が相互協調を実現できる条件を明らかにすることを目的とした．

この目的にむけて，本学位論文では，より具体的に (a.1) 強化学習のパラメータに関する条件，(a.2) 囚人のジレンマの利得行列に関する条件，(a.3) 強化学習が相互協調を実現するときの行動原理，(b) 強化学習主体が相互協調を実現する際の鍵となる性質をを明らかにすることを目標とした．これらの目標に関しては以下の知見をえた．

**(a.1) 強化学習のパラメータに関する条件**　強化学習を戦略とする繰り返し囚人のジレンマを分析した結果（3 章），強化学習戦略がともに過去の行動履歴を十分考慮して意思決定する場合ほど，強化学習戦略をとるプレイヤの間で相互協調が実現されることを示した．この結果は，一般に個人の意思決定を阻害すると思われている情報の不確実性がむしろ集団の意思決定を望ましい方向へもたらし，協調問題すなわち個人合理性と集団合理性の対立を解消できる可能性を示唆している．

**(a.2) 囚人のジレンマの利得行列に関する条件**　強化学習戦略のゲームを近似モデルで表現し分析した結果（4 章），十分に長い行動履歴を考慮する場合に相互協調解が存在する必要条件は，協調 C を選ぶことがのち

に C をだす確率を高める（$f_i(\mathrm{CC}) > 0$）ことだと示した．この必要条件から，負の利得では相互協調が実現されないことが予想されるが，この予想の正しさは本来のモデルに近い分析（3 章）により検証された．また，近似モデルのもと，強化学習が相互協調を実現しやすい利得行列を導出した．これは強化学習主体の相互協調を促進するメカニズムと位置づけできる．

**(a.3) 強化学習が相互協調を実現するときの行動原理**　相互協調を実現するときの強化学習戦略の行動パタンの分析から（3 章），相互協調を実現しているとき，強化学習戦略はしっぺ返し戦略（TFT）と類似した振る舞いを学習の結果として習得していることを示した．この結果は，学習主体による相互協調の実現方法を記述するとともに，しっぺ返し戦略が不確実な状況下での利己的学習から生じうるという位置づけを与える．

**(b) 強化学習主体が相互協調を実現する際の鍵となる性質**　協調問題の解決に関する先行研究の知見（2 章）と本論文の相互協調の条件に関する知見 (a.1)，(a.2)，および学習主体の行動原理に関する知見 (a.3) を対比させた（6 章）．強化学習と素朴定理の対比から，（長期的な将来の推論よりも）長期的な過去の経験を重視することが相互協調の実現の要因であることが示唆される．強化学習と信念学習の対比から，（利得行列や相手の行動の可視性を前提とした学習よりも）自分のアクションに対するフィードバック（個人的な利得）に基づき試行錯誤的に学習することが相互協調の実現の要因であることが示唆される．本学位論文では，以上の対比から，利得行列（利害関係）に関して不確実な状況で，長期的な過去の経験から，長期的な自己利益を最大化するという計算目標を達成することが，強化学習主体が相互協調を実現する際の鍵となる性質であることを論じた．

## 7.2　不確実な状況における学習主体の相互協調

本論文の結果は，各主体が利己的な行動原理に従い行動を学習した帰結として，集団としては協調的な振る舞いが創発することを示している．現実の社会的状況では，複数の問題が絡みあい，個人が問題の全体像を把握できることは少ない．学習とは，こうした情報の限定性から生じる不確実性に対処するため，経験に基づき行動を変化させる仕組みのひと

つである．複数の学習主体が相互作用する状況では（しばしば未知なる）相手よりもよりよく学習することが高い利益をもたらすが，学習の出来を決める要因のひとつは処理できる情報の量（あるいは情報処理能力）である．本学位論文では，主体の処理できる情報の量を「記憶容量」パラメータにより制御した．本論文の結果（3 章）によれば，不確実な状況における学習主体の相互協調は同じレベルの情報処理能力をもつ個体の間で実現可能となる．他方，相対的に相手よりも多くの情報を処理できる主体がより多くの利益をえる傾向にある．これと対照的に，問題の全体像を把握し，情報が完全で不確実性のない学習主体の協調問題は信念学習など古典的ゲーム理論や，利得情報を使った強化学習などで扱われ，その場合には相互協調は実現されない（2 章，6 章）．本論文では，相手プレイヤと利得行列に関する情報を知らない状況を考えたが，先行研究との比較から，このうち利得行列を知らないことが今回の結果により貢献していると思われる．以上は，現実で観察される協調の一部は，問題状況の複雑さと不確実性に対処する学習能力のおかげで成立している可能性を示唆している．本論文では極端に不確実な状況のみを分析対象としたが，情報の限定を緩め，相手の行動を互いに学習し合う場合（遊びなど）や，一方の個体が相手の行動パタンを予測できる場合（親子など）でもなお協調するといった，より現実的な状況での相互協調のあり方がどのように可能となるかなど，協調問題のさまざまな側面が残されている．

## 7.3　今後の展望

本学位論文では，相互協調を実現する学習主体の計算論のレベルの記述に向けて，強化学習という具体的なアルゴリズムのもと，相互協調が実現される条件を明らかにしてきた．本学位論文では，個別学習主体の計算目標やその条件に関してのみ論じており，２人ゲームの全体像すなわち両プレイヤがどのような計算目標を「同時に」達成しているのかを明らかにできたとはいえない．実際，囚人のジレンマは２人のプレイヤが同時に関与するゲームであり，その全体像は個別主体の独立した計算目標からは完全には記述できない．換言すれば，相互協調を計算論のレベルで記述するには，両プレイヤの計算目標がどのように関係しているかを理解する必要がある．また，本学位論文では先行研究の知見と理論的な扱いやすさから強化学習を用いたが，より一般的に，計算論レベルの説明を与えるには強化学習など，個別具体のアルゴリズムに依存しな

いカタチで計算目標を示す必要がある．このためには，本論文で論じた協調実現の鍵となる性質をもったより単純なモデルを分析したり，強化学習を $K$ 次戦略と捉え $K$ 次戦略と $K+1$ 次戦略との関係を分析する（5 章）など，新しい視点からの分析が必要となるだろう．

# A　囚人のジレンマの一般型

囚人のジレンマの利害構造はその拡張版にも継承されており，拡張版が扱う問題の根幹をなす．本節では２種類の拡張版ゲーム，共有地の悲劇，公共財ゲーム，を取り上げ，その定義および囚人のジレンマのジレンマ構造を導きだす．また，パラメータの範囲は限定的だが，３人ゲームの数値計算を行い，強化学習戦略が集団最適な相互協調解へ到達可能な場合があることを示す．

## A.1　共有地の悲劇

$N$ 人の集団において，$N$ 人のうち何人が協調したかによって個々人へ分配される利益が変化するゲームを考えられる．この $N$ 人版囚人のジレンマは共有地の悲劇と呼ばれる．共有地の悲劇は共有資源分配の問題を抽象化し，このゲームでは，集団が共有資源を過剰に搾取すれば枯渇し，その本来の恩恵は失われてしまう．

共有地の悲劇は囚人のジレンマの問題を根幹にもつ．形式的には，プレイヤ $i$ は「資源を搾取しない」という選択から生じる損失 $c_i \geq 0$ を被る見返りとして，資源の再生から生じる恩恵 $b_i \geq 0$ をえられる．ここで，$c_i = 0$ は搾取（損失を被らない）を表し，$b_i = 0$ は恩恵なしを表す．ゼロでない恩恵を受けるには一定人数以上の協調が不可欠である．

$N$ 人の集団で，資源を搾取しない（C）か搾取する（D）という選択肢を考え，それぞれの損失を定義する：

$$c_i = \begin{cases} c & \text{if C} \\ 0 & \text{if D} \end{cases}$$

また，C を選んだ人数を $\ell$ として，個人が受けとる恩恵を定義する：

$$b_i = \begin{cases} b & \text{if } \ell \geq \theta \\ 0 & \text{if } \ell < \theta \end{cases}$$

ここで，$\theta$ は閾値であり，$b > c > 0$ である．

もし $\theta < N-1$ であれば，損失なしに恩恵をうける機会が存在する．これはタダ乗りと呼ばれる問題である．共有地の悲劇は，自分以外の C を選んだ人数を $n$ として，次の利得表で定義される．

$$\begin{array}{c} \\ \mathrm{C} \\ \mathrm{D} \end{array} \begin{array}{c} \begin{array}{ccc} n>\theta & n=\theta & n<\theta \end{array} \\ \begin{pmatrix} b-c & b-c & -c \\ b & 0 & 0 \end{pmatrix} \end{array}$$

$N = 2$ のとき，$n \in \{0, 1\}$ より，$f_1(\mathrm{CC}) := b - c$，$f_1(\mathrm{CD}) := -c$，$f_1(\mathrm{DC}) := b$，$f_1(\mathrm{DD}) := 0$ となる．$\theta = 0.5$ では，囚人のジレンマの条件

$$f_i(\mathrm{DC}) > f_i(\mathrm{CC}) > f_i(\mathrm{DD}) > f_i(\mathrm{CD})$$

をみたす．また，$b > c$ ゆえに，繰り返し囚人のジレンマの条件

$$f_i(\mathrm{CC}) + f_i(\mathrm{CC}) > f_i(\mathrm{CD}) + f_i(\mathrm{DC})$$

をみたす．

### A.1.1 強化学習戦略と共有地の悲劇

強化学習戦略の共有地の悲劇ゲームに対する振る舞いを調べる．本分析では 3 章の分析方法を用いた．強化学習戦略のパラメータは $K = 7$，$\alpha_i = 0.8$，$\beta_i = 1$ とした（すべてのプレイヤで共通とする）．共有地の悲劇ゲームのパラメータは，人数 $N = 3$，損失 $c = 1$ に固定し，恩恵 $b = 1.5, 1.6, \ldots, 4.0$ を変化させたときの定常分布を調べた．

定常分布 $\pi$ は $|\{\mathrm{C}, \mathrm{D}\}^N| = 2^3 = 8$ 状態あるが，本論文では協調 C に関心があるため，これを C を選んだプレイヤの人数ごとに分類して調べる．具体的には，定常分布において，$N = 3$ 人中 3 人が C を選んだ確率を $\pi(3) := \pi(\mathrm{CCC})$，2 人が C を選んだ確率を $\pi(2) := \pi(\mathrm{CCD}) + \pi(\mathrm{CDC}) + \pi(\mathrm{CCD})$，1 人が C を選んだ確率を $\pi(1) := \pi(\mathrm{CDD}) + \pi(\mathrm{DCD}) + \pi(\mathrm{DDC})$，0 人が C を選んだ確率を $\pi(0) := \pi(\mathrm{DDD})$ とし，恩恵 $b$ によって $\pi(\ell)$ がどう変化するかを調べる．

タダ乗り問題が存在する条件のうち，$\theta = 1$ と $\theta = 1.5$ を図 A.1 および図 A.2 にした．$\theta = 1$ では，恩恵 $b$ が大きくなるにつれて $\pi(2)$ すなわち 2 人が協調し 1 人が裏切る状態が大半を占めるようになる．利得行列 $f$ のう

ち $\pi(2)$ に対応する要素は $f(\mathrm{C}, n=\theta) = b-c$ および $f(\mathrm{D}, n>\theta)$ であり，定常分布はこれらの 2 状態を繰り返すパタンと解釈できる．実際, $\theta = 1$ の場合, $3\,f(\mathrm{C}, n>\theta) = 3\,(b-c) < 2\,(b-c)+b = 2\,f(\mathrm{C}, n=\theta)+f(\mathrm{D}, n>\theta)$ より [1]，2 人が協調し 1 人が裏切る状態が Pareto 効率的となっている．他方，$\theta = 1.5$ では，恩恵 $b$ が大きくなるにつれて $\pi(3)$ すなわち $N=3$ 人が協調する状態が大半を占めるようになる．$\theta = 1.5$ では利得行列の中央列が消えるため，$\theta = 1$ のときとは異なり，$3\,f(\mathrm{C}, n>\theta)$ が最大の利益を与え，3 人が協調する状態が集団最適となっている．いずれの $\theta$ でも，恩恵 $b$ が小さい場合には集団最適解が最も高い確率ではないが，これは強化学習が数値としての累積利得を用いるため，$K=7$ かつ $\alpha_i = 0.8$ では十分な累積量をえられず，学習しにくいためだと考えられる．

以上から，強化学習戦略は囚人のジレンマの拡張である共有地の悲劇においても，Nash 均衡解ではなく，Pareto 効率解へ可能な場合があることを示している．共有地の悲劇はパラメータ $\theta$ によっては，十分に大きな恩恵 $b$ でも全員が協調する状態が集団最適とはならないが，その場合，強化学習戦略の集団は集団最適な解を見つける．換言すれば，学習可能な状況においては，強化学習戦略は個人最適と集団最適の対立という意味での協調問題を解決できる．

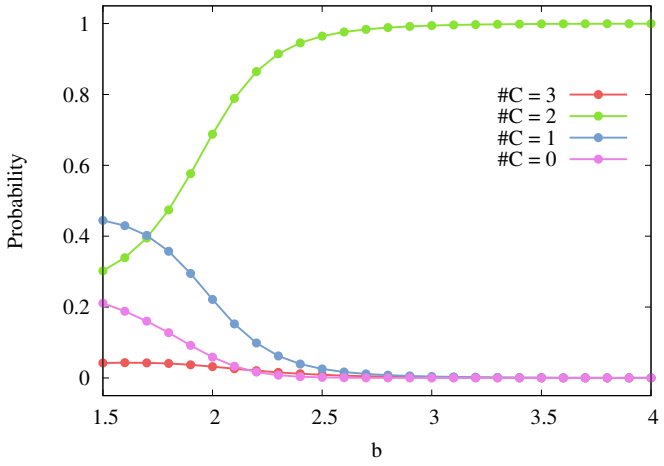



図 A.1: 強化学習戦略と共有地の悲劇（$N=3$, $\theta=1$）．協調 C を選択した人数の生起確率．$\#\mathrm{C}=3$ は全員協調，$\#\mathrm{C}=0$ は全員裏切

[1] 今回の設定のように，プレイヤ対称の設定下では全プレイヤが平均的に等しい利得をえるため，集団の合計利得を考察している．

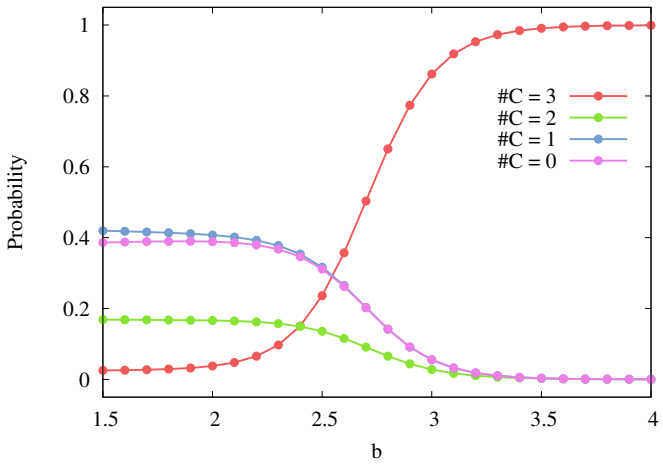



図 A.2: 強化学習戦略と共有地の悲劇（$N = 3$，$\theta = 1.5$）．協調 C を選択した人数の生起確率．$\#\mathrm{C} = 3$ は全員協調，$\#\mathrm{C} = 0$ は全員裏切

## A.2　公共財ゲーム

$N$ 人の集団において，個人がある量の費用を投資すれば，その費用の合計に比例して個々人へ等分配される利益が変化するとしよう．このゲームは公共財の性質を反映しているので，公共財ゲームと呼ばれる．公共財ゲームでは，他人が十分な投資をすれば，個人は必ずしも投資せずとも利益をえられるが，多くの個人が投資を止めてしまえば利益も失われてしまう．これはタダ乗りと呼ばれる問題である．

公共財ゲームは囚人のジレンマの問題を根幹にもつ．形式的には，プレイヤ $i$ は費用 $c_i \geq 0$ を投資した結果として，利益 $g(\sum_j c_j) - c_i$ をえる．ここで，$c_i = 0$ は投資なしを表し，また $g(x)$ は投資の効果を決める関数であり，単純な場合，$g(x) = a\,x/N$ すなわち合計費用 $x$ の定数 $a$ 倍だけ利益が生じ，それを全員で等分配する．

2 人の集団で，投資する（C）か投資しない（D）しか選択肢がない場合を考え，それぞれの費用を定義する：

$$c_i = \begin{cases} c & \text{if C} \\ 0 & \text{if D} \end{cases}$$

両者とも投資するとき $f_i(\mathrm{CC}) := a\,c - c$, 自分のみ投資するとき $f_i(\mathrm{CD}) := a\,c/2 - c$，相手のみ投資するとき $f_i(\mathrm{DC}) := a\,c/2$，両者とも投資しな

いとき $f_i(\mathrm{DD}) := 0$ となる．公共財ゲームがタダ乗り問題を含むのは $f_i(\mathrm{CC}) > f_i(\mathrm{DD})$ かつ $f_i(\mathrm{CD}) < f_i(\mathrm{DD})$ のときであるが，ここから $1 < a < 2$ が導かれる．また $a < 2$ の範囲で $f_i(\mathrm{DC}) > f_i(\mathrm{CC})$ であるから，囚人のジレンマの条件

$$f_i(\mathrm{DC}) > f_i(\mathrm{CC}) > f_i(\mathrm{DD}) > f_i(\mathrm{CD})$$

をみたす．さらに $a > 1$ の範囲で繰り返し囚人のジレンマの条件

$$f_i(\mathrm{CC}) + f_i(\mathrm{CC}) > f_i(\mathrm{CD}) + f_i(\mathrm{DC})$$

をみたす．したがって，囚人のジレンマと本質的に類似したジレンマ構造をもつ．プレイヤ数 $N > 2$ では，$N$ 人中何人が投資すれば正の利益が生じるかによって倍率 $a$ は異なるが，公共財ゲームの条件から $1 < a < N$ が導かれる．

### A.2.1 強化学習戦略と公共財ゲーム

強化学習戦略の公共財ゲームに対する振る舞いを調べる．本分析では第 3 章の分析方法を用いた．強化学習戦略のパラメータは $K = 7$, $\alpha_i = 0.8$, $\beta_i = 4$ とした（すべてのプレイヤで共通とする）[2]．公共財ゲームのパラメータは，人数 $N = 3$，損失 $c = 1$ に固定し，倍率 $a = 1.05, 1.1, \ldots, 2.0$ を変化させたときの定常分布を調べた．

定常分布 $\pi$ は $|\{\mathrm{C}, \mathrm{D}\}^N| = 2^3 = 8$ 状態あるが，本論文では協調 C に関心があるため，これを C を選んだプレイヤの人数ごとに分類して調べる．共有地の悲劇の分析と同様に，定常分布において，$N = 3$ 人中 3 人が C を選んだ確率を $\pi(3) := \pi(\mathrm{CCC})$，2 人が C を選んだ確率を $\pi(2) := \pi(\mathrm{CCD}) + \pi(\mathrm{CDC}) + \pi(\mathrm{CCD})$，1 人が C を選んだ確率を $\pi(1) := \pi(\mathrm{CDD}) + \pi(\mathrm{DCD}) + \pi(\mathrm{DDC})$，0 人が C を選んだ確率を $\pi(0) := \pi(\mathrm{DDD})$ とし，倍率 $a$ によって $\pi(\ell)$ がどう変化するかを調べる．

図 A.3 から，倍率 $a$ が大きくなるにつれて $\pi(3)$ すなわち $N = 3$ 人が協調する状態が大半を占めるようになる．線形の $g(x) = ax/N$ を用いる場合，C を選んだ人数を $\ell$ とするとき，集団の合計利得は $\ell\,(a-1)\,c > 0$ となる．したがって，$a$ に依らず，3 人が協調する状態が集団最適となる．また，倍率 $1 < a < N$ が大きくなるほど，協調する誘因は高まるが，こ

[2] $\beta_i = 4$ としたのは，利得の絶対値が小さいため．

れは図 A.3 と一致する．倍率 $a$ が小さい場合には 3 人が協調する状態が最も高い確率ではないが，共有地の悲劇の場合と同じく，これは強化学習が数値としての累積利得を用いるため，$K = 7$ かつ $\alpha_i = 0.8$ では十分な累積量をえられず，学習しにくいためだと考えられる．実際，$a > 1$ がほとんど 1 に近いとき，集団の合計利得は $\ell\,(a-1)\,c \approx 0$ となり[3]，全員裏切 D の利得との差分が小さい．

以上から，強化学習戦略は囚人のジレンマの拡張である公共財ゲームにおいても，Nash 均衡解ではなく，Pareto 効率解へ到達可能な場合があることを示している．換言すれば，学習可能な状況においては，強化学習戦略は個人最適と集団最適の対立という意味での協調問題を解決できる．

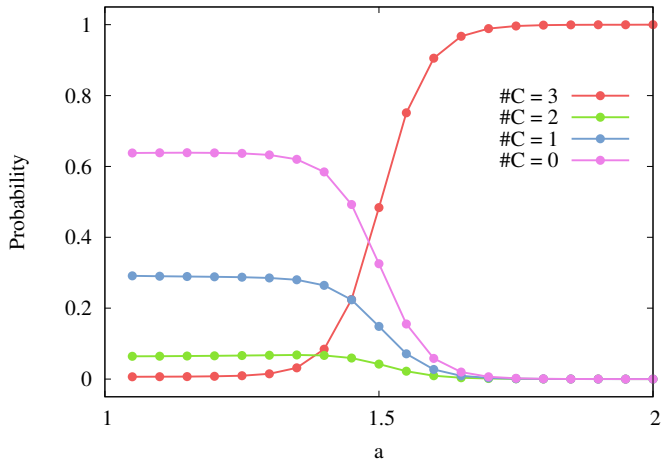



図 A.3: 強化学習戦略と公共財ゲーム（$N = 3$，$g(x) = ax/N$）．協調 C を選択した人数の生起確率．$\#\mathrm{C} = 3$ は全員協調，$\#\mathrm{C} = 0$ は全員裏切

[3]共有地の悲劇の場合と同じく，プレイヤ対称の設定下では全プレイヤが平均的に等しい利得をえるため，集団の合計利得を考察している．

# 謝辞

本学位論文は，橋本敬教授の幅広い関心と寛容なご指導のおかげで完成できたと思われます．本学位論文のテーマは，修士課程におけるコミュニケーションに関するテーマから一転し，博士課程の途中から取り組んだテーマでした．橋本教授は本テーマで学位論文を提出することを寛容にも受け入れてくださいました．ここにできる限りの努力のもとで一応の完成に至るまで見守ってくださったことに感謝いたします．

本学位論文のテーマは，日高昇平助教および真隅暁さんとの交流をきっかけに始まりました．本学位論文のテーマは実質的にお二方との議論のなかで進めてきたといえます．とりわけ日高助教には問題の捉え方，数学的・技術的な考え方に関して多くの助言と刺激を頂きました．辛抱強くご指導いただいた内容は本学位論文の執筆にあたり不可欠だったと思います．「問題はその先にある」という助言はこころに刻んでおきます．

本学位論文の完成に必要なスキルは，橋本教授と日高助教のお二方のご指導からえられたものであることは疑いようがありません．ここで，お二方の異なる考え方に触れることができたのは，スキルの獲得に繋がるとともに，何事にも変えがたい経験だったと感じております．これから歩む先にも，ときどき道を照らしていただければと思います．

本学位論文の審査においては，外部審査員の野田五十樹先生をはじめ，審査員の中森義輝先生，ヒュン・ナム・ヤン先生，ダム・ヒョウ・チ先生には，建設的な議論を展開していただき，感謝いたします．本学位論文の草稿の多くの不備が修正され，洗練されたものを提出することができました．野田先生からは人工知能学会の質疑でも助言をいただき，その後の研究の方向性に反映されました．また，草稿を読み，専門的な観点から助言をくださった佐々木康朗先生に感謝いたします．

本学位論文の提出にあたっては，現在私が東京で勤務している都合上，橋本研究室の田村香織さんには，論文の製本や書類の提出など手助けをいただきました．最後に，遠方から支援してくれた父母妹弟，ともに学んできた研究室の皆さまに，あわせて感謝いたします．　　鳥居拓馬

# 参考文献


[1] J. McKenzie Alexander. Cooperation. In Sahorta Sarkar and Anya Plutynski, editors, *Companion to the Philosophy of Biology*, chapter 22, pages 415–430. Blackwell Publishing: Oxford, 2008.

[2] James Andreoni and John H. Miller. Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence. *The Economic Journal*, 103(418):570–585, 1993.

[3] Robert Axelrod. Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, 24(1):3–25, 1980.

[4] Robert Axelrod. The emergence of cooperation among egoists. *American Political Science Review*, 75(2):306–318, 1981.

[5] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[6] Dipyaman Banerjee and Sandip Sen. Reaching pareto-optimality in prisoner's dilemma using conditional joint action learning. *Autonomous Agents and Multi-Agent Systems*, 15(1):91–108, 2007.

[7] Pedro Dal Bo. Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *The American Economic Review*, 95(5):1591–1601, 2005.

[8] Tilman Borgers and Rajiv Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77:1–14, 1997.

[9] Tilman Borgers and Rajiv Sarin. Naive reinforcement learning with endogenous aspirations. *International Economic Review*, 41(4):921–950, 2000.

[10] Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

[11] Steven J. Brams. Newcomb's problem and prisoners' dilemma. *The Journal of Conflict Resolution*, 19(4):596–612, 1975.

[12] George W. Brown. Iterative solution of games by fictitious play. In T.C. Koopmans, editor, *Activity Analysis of Production and Allocation*, pages 374–376. New York: John Wiley & Sons, 1951.

[13] Robert R. Bush and Frederick Mosteller. A mathematical model for simple learning. *Psychological Review*, 58:413–423, 1951.

[14] Robert R. Bush and Frederick Mosteller. A stochastic model with application to learning. *The Annals of Mathematical Statistics*, 24 (4):559–585, 1953.

[15] Angel Cabrera and Elizabeth F. Cabrera. Knowledge-sharing dilemmas. *Organization Studies*, 23(5), 2002.

[16] Colin F. Camerer. Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences*, 7(5):225–231, 2003.

[17] Colin F. Camerer and Teck Hua Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999.

[18] Gerald Carter. The reciprocity controversy. *Animal Behavior and Cognition*, 1(3):368–386, 2014.

[19] Russell Cooper, Douglas V. De Jong, Robert Forsythe, and Thomas W. Ross. Cooperation without reputation: experimental evidence from prisoner's dilemma games. *Games and Economic Behavior*, 12:187–218, 1996.

[20] John G. Cross. A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, 87(2):239–266, 1973.

[21] Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.

[22] Andreas Flache and Michael W. Macy. Stochastic collusion and the power law of learning: a general reinforcement learning model of cooperation. *The Journal of Conflict Resolution*, 46(5):629–653, 2002.

[23] James W. Friedman. A non-cooperative equilibrium for supergames. *The Review of Economic Studies*, 38(1):1–12, 1971.

[24] William D. Hamilton. The evolution of altruistic behavior. *The American Naturalist*, 97(896):354–356, 1963.

[25] William D. Hamilton. The genetical evolution of social behavior. *Journal of Theoretical Biology*, 7:1–16, 1964.

[26] Shohei Hidaka, Takuma Torii, and Akira Masumi. Which types of learning make a simple game complex? *Complex Systems*, 24(1): 49–74, 2015.

[27] Nigel Howard. *Paradoxes of Rationality: Theory of Metagames and Political Behavior*. Cambridge: MIT Press, 1971.

[28] Lawrence M. Kahn and J. Keith Murnighan. Conjecture, uncertainty, and cooperation in prisoner's dilemma games. *Journal of Economic Behavior and Organization*, 22:91–117, 1993.

[29] Lawrence M. Kahn and J. Keith Murnighan. *Payoff uncertainty and cooperation in finitely-repeated prisoner's dilemma games*, chapter 65. Elsevier, 2008.

[30] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decisions under risk. *Econometrica*, 47(2):263–291, 1979.

[31] Ardeshir Kianercy and Aram Galstyan. Dynamics of boltzmann q-learning in two-player two-action games. *Physical Review E*, 85: 041145–1–9, 2012.

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

[33] David M. Kreps, Paul Milgrom, John Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoner's dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

[34] Michael W. Macy. Learning to cooperate: stochastic and tacit collusion in social exchange. *American Journal of Sociology*, 97(3): 808–843, 1991.

[35] David Marr. *Vision*. Cambridge: MIT Press, 1982.

[36] Naoki Masuda and Mitsuhiro Nakamura. Numerical analysis of a reinforcement learning model with the dynamic aspiration level in the iterated prisoner's dilemma. *Journal of Theoretical Biology*, 278: 55–62, 2011.

[37] Naoki Masuda and Hisashi Ohtsuki. A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. *Bulletin of Mathematical Biology*, 71:1818–1850, 2009.

[38] Manfred Milinski and Claus Wedekind. Working memory constrains human cooperation in the prisoner's dilemma. *Proceedings of the National Academy of Sciences*, 95:13755–13758, 1998.

[39] Per Molender. The optimal level of generosity in a selfish, uncertain environment. *Journal of Conflict Resolution*, 29(4):611–618, 1985.

[40] Claire El Mouden, Maxwell Burton-Chellew, Andy Gardner, and Stuart A. West. What do humans maximize? In Samir Okasha and Ken Binmore, editors, *Evolution and Rationality: Decisions, Co-operation and Strategic Behaviour*, chapter 2, pages 23–49. Cambridge University Press, 2012.

[41] Martin A. Nowak. Stochastic strategies in the prisoner's dilemma. *Theoretical Population Biology*, 38:93–112, 1990.

[42] Martin A. Nowak. *Evolutionary Dynamics*. The Belknap Press of Harvard University Press, 2006.

[43] Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 134:1560–1563, 2006.

[44] Martin A. Nowak. Evolving cooperation. *Journal of Evolutionary Biology*, 299:1–8, 2012.

[45] Martin A. Nowak and Karl Sigmund. Tit-for-tat in heterogeneous populations. *Nature*, 355:250–253, 1992.

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

[47] William H. Press and Freeman J. Dyson. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, 109(26):10409–10413, 2012.

[48] Anatol Rapoport and Albert M. Chammah. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press, 1965.

[49] Alvin E. Roth and Ido Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8:164–212, 1995.

[50] Tuomas W. Sandholm and Robert H. Crites. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37 (1–2):147–166, 1996.

[51] Yuzuru Sato and James P. Crutchfield. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1):015206–015210, 2003.

[52] Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002.

[53] Reinhard Selten and Rolf Stoecker. End behavior in sequences of finite prisoner's dilemma supergames: a learning theory approach. *Journal of Economic Behavior and Organization*, 7:47–70, 1986.

[54] Herbert A. Simon. *Administrative Behavior*. Macmillan, 1947.

[55] Satinder P. Singh, Michael J. Kearns, and Yishay Mansour. Nash convergence of gradient dynamics in general-sum games. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI'00)*, pages 541–548, 2000.

[56] Brian Skyrms. *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, 2003.

[57] Alexander J. Stewart and Joshua B. Plotkin. Extortion and cooperation in the prisoner's dilemma. *Proceedings of the National Academy of Sciences*, 109(26):10134–10135, 2012.

[58] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[59] Karl Tuyls, Pieter Jan'T Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, 12:115–153, 2006.

[60] Marcel van Assen, Chris Snijders, and Jeroen Weesie. Behavior in repeated prisoner's dilemma games with shifted outcomes analyzed with a statistical learning model. *Journal of Mathematical Sociology*, 30:159–180, 2006.

[61] Marcel A.L. van Assen and Chris Snijders. Effects of risk preferences in social dilemmas: a game-theoretical analysis and evidence from two experiments. In Ramzi Suleiman, David V. Budescu, Ilan Fischer, and David M. Messick, editors, *Contemporary Psychological Research on Social Dilemmas*. Cambridge University Press, 2004.

[62] Paul A.M. van Lange, Jeff Joireman, Craig D. Parks, and Eric van Dijk. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, 120:125–141, 2013.

[63] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[64] Stuart A. West, Ashleigh S. Griffin, and Andy Gardner. Social semantics: altruism, cooperation, mutualism, strong reciprocity and

group selection. *Journal of Evolutionary Biology*, 20(2):415–432, 2007.

[65] Wako Yoshida, Ray J. Dolan, and Karl J. Friston. Game theory of mind. *PLoS Computational Biology*, 4(12):e1000254 1–14, 2008.

[66] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of The Twentieth International Conference on Machine Learning (ICML 2003)*, pages 1–8, 2003.

[67] 酒井 泰弘. フランク・ナイトの経済思想 —リスクと不確実性の概念を中心として—. Discussion Paper J-19, 滋賀大学経済学部附属リスク研究センター, 2012.

[68] 鳥居 拓馬, 日高 昇平, and 真隅 暁. 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. In **第 *28* 回 人工知能学会全国大会予稿集**, 2014. 4H1-3.