様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２７ 日現在

機関番号：１２６１２
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２０７５２

研究課題名（和文）不完全情報下での逐次的意思決定：部分観測マルコフ決定過程解法の探索

研究課題名（英文）Sequential Decision Making with Imperfect Information: An application of POMDP

研究代表者
岩崎　敦（Iwasaki, Atsushi）

電気通信大学・大学院情報理工学研究科・准教授

研究者番号：３０３８０６７９

交付決定額（研究期間全体）：（直接経費）　　5,000,000円

研究成果の概要（和文）：本研究では，不完全情報下における逐次的意思決定の分析手法を開拓することを目的とする．具体的には，私的観測というお互いの行動を正確に観測できない不完全観測下において繰り返し行われる意思決定をくり返しゲームの枠組みで考え，その帰結（均衡）を求める問題を扱う．まず，進化ゲームでよく用いられる突然変異付きレプリケータダイナミクスを利用した分析を進め，有名なしっぺ返し戦略の代わりに勝ち残り・負け逃げ戦略が優位になる条件を明らかにした．次に，そのダイナミクスの構造を利用した正則化先導者追従（FTRL）ベースの均衡計算アルゴリズムを開発し、N人単調ゲームを扱えることを証明した．

研究成果の学術的意義や社会的意義
人がどのように協力する／しないかの仕組みは学際的な研究課題であり，繰り返しゲームは，いつ終わるかわからない相手との関係が協力を誘発するとして，その仕組みを解明する研究分野である．その中でも私的観測は，その有用性を指摘されながらも明らかになっていないことが多い研究課題である。これに対して本研究は、進化ゲームの枠組みを利用して、行動の取り違えにおける新しい戦略である単独裏切-相互処罰戦略を発見した。さらにその枠組みを学習アルゴリズムに応用し、私的観測のようなノイズ下でも均衡を計算できるアルゴリズムを開発した。

研究成果の概要（英文）：This work aims to develop an analytical method for sequential decision-making under imperfect information. Specifically, we utilize a repeated game framework under private monitoring, where each player cannot directly observe the actions of others. We seek to determine the outcome (equilibrium) of such decision-making processes. First, we analyze the problem using the replicator-mutator dynamics commonly used in evolutionary games and identify the conditions under which Tit-For-Tat is replaced by Win-Stay, Lose-Shift. Next, we develop a mutation-driven Follow-The-Regularized-Leader (FTRL) algorithm based on the structure of this dynamics, and prove it handle N-player monotone games, which incluedes two-player zero-sum games and Cournot competitions.

研究分野：ゲーム理論

キーワード：ゲーム理論　繰り返しゲーム　アルゴリズム　最適化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

長期的関係にあるプレイヤが、いかにして協調関係を維持するかを解明することはゲーム理論における中心的な研究課題の一つである。プレイヤがお互いの過去の行動を完全に観測できる完全観測下の無限回の繰り返しゲームにおいては、将来の利得を十分大きく見積もる（割引因子が 1 に近い）とき、任意の協力関係が維持可能であることを示すフォーク定理等の理論的成果がある。それに対して、より現実的な不完全観測では観測にノイズが伴う (Mailath and Samuelson 2006)。不完全観測は、お互いの行動に関する不完全な情報を全員が共有する「公的観測」と、その不完全な情報を全く共有しない「私的観測」に分かれる。公的観測は、完全観測とほぼ同様な結果が示されている。これに対して私的観測下では、プレイヤの観測がお互いに異なる、つまりプレイヤはお互いの行動に関する共通の情報をもたないため、相手の行動を予測することが極めて困難になる。例えばあるプレイヤは、自分の持つ情報、つまり、自分の行動履歴、自分の（相手の行動に関する）観測履歴、そして、自分の行動から生じる相手の観測履歴に関する予想（信念）から、次の自分の行動を決定する。このため可能な履歴の数は、ゲームの繰り返し数に対して指数的に増加する。つまり、完全観測と比べ、私的観測下では、ノイズを含む観測に対して複雑な統計的推論を必要とする。このため、私的観測という仮定は極めて自然であり様々な応用事例が表現できるにも関わらず、私的観測下での繰り返しゲームの解析は極めてチャレンジングな課題として残っている。応募者は私的観測の問題を POMDP 問題に帰着して効率的に解く手法を模索してきた (Joe et al. 2012)。 しかし現状では、近似解の定義すら困難であり、これまでと全く異なるパラダイムの定式化や解法が必要であると考えた。そこで諸分野の似た問題や数理技術を調査してきた結果、計算機科学の諸分野に可能性を感じ、本研究の構想を得た。

Mailath and Samuelson. Repeated Games and Reputations: Long-run Relationships. Oxford Univ. Press, 2006.
Joe, Iwasaki, Kandori, Obara, and Yokoo. Automated Equilibrium Analysis of Repeated Games with Private Monitoring: A POMDP Approach. Proceedings of the 11th International Joint Conference on Autonomous Agents and Multiagent Systems, 1305–1306, 2012.

２．研究の目的

本研究では、計算機科学の諸分野の理論から、不完全情報下における逐次的意思決定の分析手法を開拓することを目的とする。具体的には、私的観測というお互いの行動を正確に観測できない不完全観測下において繰り返し行われる意思決定を繰り返しゲーム理論の枠組みで考え、そのゲームの帰結 (均衡) を求める問題を扱う。これまで、不完全観測、とくに私的観測と呼ばれる環境下において、繰り返しゲームの広範的な解析は行われてこなかった。この状況では、プレイヤは相手の行動を予測するため、自らのノイズを含む観測をもとに、相手の観測履歴を統計的に推論しなければならない。このため推論対象となる観測履歴の数はゲームの繰り返し数に対して指数的に増加する。これは部分観測マルコフ決定過程 (Partially Observable Markov Decision process, POMDP) (Sondik 1978) に帰着できることが知られているが、一般には決定不能 (UNDECIDABLE) な問題であり、解析的な分析が可能な定式化や解法は未だ見つかっていない。そこで、近年発展が著しい機械学習理論／制御理論／情報理論といった諸分野の理論から POMDP 問題を俯瞰し、大規模な問題に適用可能な、精度保証つきの近似解法を構築する。

Sondik. The optimal control of partially observable Markov processes over the infinite horizon: discounted cost. Operations Research, 26(2):282–304, 1978.

３．研究の方法

はじめに、囚人のジレンマおよびその一般化であるクールノー競争を対象に、機械学習理論を用いた近似 POMDP アルゴリズムを開発する。とくに、確率的勾配降下法や後悔最小化学習を検討する。とくに後者の後悔最小化学習はそのダイナミクスが粗相関均衡に収束することが知られているが、不完全情報下での挙動は。テキサスホールデムポーカーといったゼロサムゲームのナッシュ均衡（Bowling et al. 2015）に限られている。本研究では、これらのアルゴリズムの成果を一般のゲームにおける不完全観測に拡張する。

次に、POMDP 問題の定式化そのものを工夫し、従来と異なる解法を試みる。具体的には、制御理論におけるフィルタリング問題に着目する。フィルタリング問題とは、機器を適切に制御するために、機器から得るセンサデータからノイズを濾し取り、機器の状態を正しく推定する問題である。本研究では、POMDP 問題をフィルタリング問題として定式化し、ノイズを含む観測に対する統計的推論を制御理論由来のアルゴリズムでどこまで可能かを明らかにする。

これら２つの課題を相互にフィードバックさせながら取り組み、大規模な問題に適用可能な精度を保証する近似解法の基盤を何らかの形で構築できると予想している。しかしその得た解が

どのような規範的な意味をもつかはわからない。そこで、近似解の構造をもとに新しい均衡概念が設計する一方、設計した均衡概念に応じて近似解を改良する。例えば、プレイヤが他の戦略に逸脱したときの利得の増加分が、ある閾値以下であることを保証する戦略の組合せとして与えられる実行可能な時間内で計算可能な新しい均衡概念を提案する。

Bowling et al. Heads-up Limit Hold'em Poker is Solved. Science, 347(6218):145–149, 2015.

４．研究成果

本研究課題は、大きく分けて２つの研究成果を挙げた。一つは突然変異付きレプリケータダイナミクスを用いた、不完全観測下の繰り返しゲームの分析であり、もう一つはそのダイナミクスを基にした、均衡計算のための学習アルゴリズムの設計と評価である。

制御理論におけるフィルタリング問題を調査した結果、微分方程式系で戦略の挙動を分析することにした。このとき、進化ゲームの文脈で用いられている数多くなるダイナミクスの中で、突然変異付きレプリケータダイナミクスに着目する。これは無限集団上で各戦略をとるプレイヤの頻度の時間的変化を、それぞれの戦略の適応度と突然変異で表現するダイナミクスである。私的観測下では、昨日までの観測と行動の履歴によって今日の行動を決める。無限回繰り返しゲームでは履歴の大きさが無限大になってしまうので、オートマトンによる簡略表現を用いることが多い。そこで、状態数 2 以下の非同相なオートマトンを列挙した戦略空間上に利得表を構築し、ダイナミクスの帰結を分析した。その結果、一般によく知られているしっぺ返し戦略（Tit-For-Tat）の代わりに勝ち残り・負け逃げ戦略（Win-Stay, Lose-Shift, 図 1）が優位になる条件を明らかにした。この成果は第 19 回情報科学技術フォーラムで発表したのち、FIT2020 船井ベストペーパー賞を受賞し、情報処理学会論文誌に掲載された。

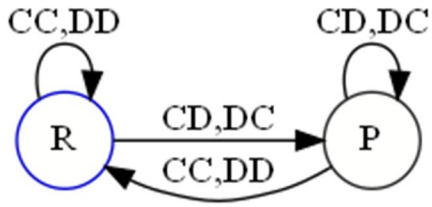

図 1：勝ち残り・負け逃げ戦略

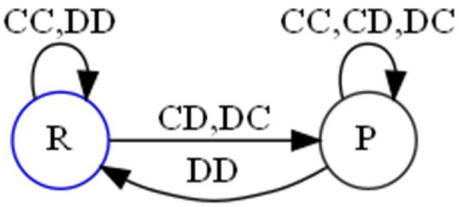

図 2：単独裏切-相互処罰戦略

理論生物学分野では、私的観測は行動の見間違え、公的観測は行動の取り違えと対応付けられており、主に行動の取り違えにおいて広範な研究が行われてきた。そこで行動の見間違えと取り違えという観測構造の違いがゲームの帰結にどんな影響を与えるかを吟味した。理論生物学における行動の取り違えは「1 期記憶戦略（Memory-one strategies）」という戦略のクラス（例えば Baek et al. 2016）をもとに解析されており、協力のコストが十分小さいとき、勝ち残り・負け逃げ戦略が生き残りやすいことが知られている。我々は、行動の見間違えにおける 2 状態オートマトン戦略の空間が、行動の取り違えにおける 1 期記憶戦略とほぼ等価であることを明らかにした。一方で、この 1 期記憶戦略にしたがうプレイヤは、昨日実現した行動の組だけを観察して、今日の行動を決めており、昨日どんな行動を意図していたかを考慮していないことがわかった。つまり、昨日お互いに協力が実現したとき、プレイヤは自分が協力しようとして協力したのか、裏切ろうとしたけど、間違えて協力したのかを区別せずに今日の行動を決定していた。

Baek et al. Comparing reactive and memory-one strategies of direct reciprocity. Sci. Rep. 6, 25676, 2016.

そこで、行動の取り違えにおいて、プレイヤの昨日の意図を考慮しない 1 期記憶戦略空間（32 戦略）を、昨日の意図を考慮する 2 状態オートマトン戦略空間（482 戦略）に拡張し、突然変異付きレプリケータダイナミクスの帰結を分析した。その結果、1 期記憶戦略は常に裏切る戦略と勝ち残り・負け逃げ戦略を除いて生き残ることはなかった。さらに、勝ち残り・負け逃げ戦略より広いパラメータの組み合わせにおいて生き残る戦略として、単独裏切-相互処罰戦略（図 2）を発見した。この戦略は裏切行動をとる状態 P での状態遷移が、勝ち残り・負け逃げ戦略とほんのわずかに異なる。お互いに協力が実現したあとの状態遷移のみが異なるだけの戦略にも関わらず、482 戦略間の相互作用の結果、勝ち残り・負け逃げ戦略を淘汰する戦略になっている。これに至る成果は国内学会で累計 9 件あり、最新の内容を第 21 回情報科学技術フォーラムで発表予定である。さらに英語論文誌への投稿を準備している。

機械学習理論を用いた近似 POMDP アルゴリズムの基礎として、均衡計算のための学習アルゴリズムの設計と解析を進めた。具体的には、突然変異付きレプリケータダイナミクスの構造を利用した正則化先導者追従（Follow the Regularized Leader, FTRL）法ベースの均衡計算アルゴリズムを開発した。まずは、Hennes et al. 2020 の Neural Replicator Dynamics を Neural Replicator-Mutator Dynamics に拡張し、囚人のジレンマやクールノー競争における均衡戦略を計算した。

この成果は第 20 回情報科学技術フォーラムで発表したのち、FIT2021 船井ベストペーパー賞を受賞した。

Hennes et al. Neural Replicator Dynamics: Multiagent Learning via Hedging Policy Gradients. In AAMAS, 492–501, 2020.

次に、2 人ゼロサムゲームに特化して、Neural Replicator-Mutator Dynamics の連続時間ダイナミクスを解析した。これはクールノー競争では、複雑すぎて、解析の糸口が見えなかったため、2 人ゼロサムゲームにおける終極反復収束（Last-iterate convergence）を Neural Replicator-Mutator Dynamics が満たすか否かを解析した。2 人ゼロサムゲームでさえも、FTRL 法ベースのアルゴリズムによるダイナミクスは、ナッシュ均衡を中心とした周回軌道をとることが知られている。この周回軌道を平均するとナッシュ均衡と一致する性質を平均反復収束（Average-iterate convergence）と言い、周回軌道に陥らず、ダイナミクスの終極が直接ナッシュ均衡に収束する性質を終極反復収束と呼ぶ（Daskalakis and Panageas 2019）。この性質を満たすよう FTRL 法に突然変異を取り入れた Mutation-driven FTRL を提案した。この内容は人工知能分野のトップ会議である UAI2022 に採択された。

Daskalakis and Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. Innovations in Theoretical Computer Science, 2019.

さらに、FTRL 法の代わりに、乗算型重み更新法に突然変異を導入した Mutant Multiplicative Weight Update (M2WU)を提案した。これは離散時間ダイナミクスでの終極反復収束を示しただけでなく、私的観測と似た不完全な（利得の）勾配情報を特定の方向へとわずかに変異させることで、終極反復収束を満たすことに成功した。この内容な人工知能分野のトップ会議である AISTATS2023 に採択された。現在、クールノー競争を含む N 人単調ゲームにおいて終極反復収束を満たす学習アルゴリズムを開発し、NeurIPS2023 や英語論文誌への投稿を進めている。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 西野上 和真、五十嵐 瞭平、岩崎 敦	４．巻 63
２．論文標題 私的観測下の繰り返し囚人のジレンマにおける協力のダイナミクス	５．発行年 2022年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1138～1148
掲載論文のDOI（デジタルオブジェクト識別子） 10.20729/00217615	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, Atsushi Iwasaki	４．巻 206
２．論文標題 Last-Iterate Convergence with Full and Noisy Feedback in Two-Player Zero-Sum Games	５．発行年 2023年
３．雑誌名 Proceedings of The 26th International Conference on Artificial Intelligence and Statistics	６．最初と最後の頁 7999-8028
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kenshi Abe, Mitsuki Sakamoto, Atsushi Iwasaki	４．巻 180
２．論文標題 Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games	５．発行年 2022年
３．雑誌名 Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence PMLR	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計19件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 村井 伸一郎
２．発表標題 取り違えのある繰り返し囚人のジレンマにおける単独裏切-相互同期戦略
３．学会等名 情報処理学会第85回全国大会
４．発表年 2023年

１．発表者名
村井 伸一郎

２．発表標題
取り違えのある繰り返し囚人のジレンマにおける協力のダイナミクス

３．学会等名
第２１回情報科学技術フォーラム（選奨論文）

４．発表年
2022年

１．発表者名
豊島 健太郎

２．発表標題
二人零和ゲームにおける突然変異駆動型Follow-The-Regularized-Leaderの終極反復収束

３．学会等名
第２１回情報科学技術フォーラム（選奨論文）

４．発表年
2022年

１．発表者名
村井 伸一郎

２．発表標題
取り違えのある繰り返し囚人のジレンマにおける協力のダイナミクス

３．学会等名
人工知能学会全国大会

４．発表年
2022年

１．発表者名
豊島 健太郎

２．発表標題
二人零和ゲームにおける突然変異付きレプリケータダイナミクスを用いた学習アルゴリズムに関する研究

３．学会等名
人工知能学会全国大会

４．発表年
2022年

１．発表者名
坂本充生

２．発表標題
二人零和ゲームにおける突然変異付きレプリケータダイナミクスを用いた学習アルゴリズムに関する研究

３．学会等名
情報処理学会第84回全国大会

４．発表年
2022年

１．発表者名
五十嵐瞭平

２．発表標題
ほぼ公的観測下の繰り返しプロジェクトゲームにおける協力のダイナミクス

３．学会等名
情報処理学会第84回全国大会

４．発表年
2022年

１．発表者名
豊島健太郎

２．発表標題
クールノー競争におけるマルチエージェント強化学習に関する研究

３．学会等名
情報処理学会第84回全国大会

４．発表年
2022年

１．発表者名
村井伸一郎

２．発表標題
取り違えのある繰り返し囚人のジレンマにおける協力のダイナミクス

３．学会等名
情報処理学会第84回全国大会

４．発表年
2022年

１．発表者名
坂本充生

２．発表標題
見間違えのある繰り返しゲームのためのActor-Critic型強化学習

３．学会等名
第24回情報論的学習理論ワークショップ（IBIS2021）

４．発表年
2021年

１．発表者名
坂本充生

２．発表標題
見間違えのある繰り返しゲームのためのActor-Critic型強化学習

３．学会等名
日本OR学会秋季研究発表会

４．発表年
2021年

１．発表者名
五十嵐瞭平

２．発表標題
ほぼ公的観測下の囚人のジレンマにおける協力のダイナミクス

３．学会等名
日本OR学会秋季研究発表会

４．発表年
2021年

１．発表者名
坂本充生

２．発表標題
見間違えのある繰り返し囚人のジレンマにおける方策勾配法に関する研究

３．学会等名
第２０回情報科学技術フォーラム（選奨論文）

４．発表年
2021年

１．発表者名
五十嵐瞭平

２．発表標題
ほぼ公的観測下の繰り返し囚人のジレンマにおける協力のダイナミクス

３．学会等名
第２０回情報科学技術フォーラム（選奨論文）

４．発表年
2021年

１．発表者名
五十嵐瞭平

２．発表標題
ほぼ公的観測下の繰り返し囚人のジレンマにおける協力のダイナミクス

３．学会等名
人工知能学会全国大会

４．発表年
2021年

１．発表者名
坂本充生

２．発表標題
見間違えのある繰り返し囚人のジレンマにおける()学習に関する研究

３．学会等名
人工知能学会全国大会

４．発表年
2021年

１．発表者名
島野 雄貴

２．発表標題
反実仮想後悔最小化によるアメリカンフットボールにおけるオフェンス戦略の均衡推定

３．学会等名
人工知能学会全国大会

４．発表年
2021年

１．発表者名
西野上和真

２．発表標題
私的観測下の繰り返し囚人のジレンマにおける協力のダイナミクス

３．学会等名
FIT2020第19回情報科学技術フォーラム

４．発表年
2020年

１．発表者名
岩崎敦

２．発表標題
見間違えのある繰り返し囚人のジレンマにおける協力の発生と振動

３．学会等名
日本オペレーションズ・リサーチ学会2021年春季研究シンポジウム（招待講演）

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Computational Incentive Science
https://sites.google.com/site/a2ciwasaki/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関