未知環境における多人数不完全情報ゲームの戦略計算

河村 圭悟[1,a)]　鈴木潤[2,3,b)]　鶴岡 慶雅[4]

概要：近年の不完全情報ゲームの分野においては，counterfactual regret minimization やその考え方を応用した DeepStack など，ゲームの状態遷移規則を事前知識として用い，ゲームの木を探索することで大規模なゲームにおける強い戦略プロファイルを求める手法が発展してきている．その一方，強化学習という枠組みにおいては，Markov decision process で表せる 1 人ゲームに対して，ゲームの状態遷移規則などが未知の状態から環境との相互作用を通じて学習し，問題を解くという学習方式がよく用いられる．
本研究では，多人数の不完全情報ゲームを，強化学習のようにゲームの状態遷移規則を用いることなく解くという学習方式を提案し，定式化する．この学習方式でゲームを効果的に解く方法論を考案することで，未知の環境であっても，多人数ゲームに帰着できる問題であることがわかれば，プレイすることで最適な戦略を得ることができる．本論文では，この枠組みに対していくつかの既存手法を適用し，最適戦略に対する評価を行うことで結果を比較した．本研究の貢献として，多人数不完全情報ゲームについてゲームの状態遷移規則を事前に用いることなくナッシュ均衡戦略を求める学習方式を，我々の知る限り初めて定式化したことが挙げられる．

Computing Strategies for Multi-player Imperfect-Information Games in Unknown Environments

Keigo Kawamura[1,a)]　Jun Suzuki[2,3,b)]　Yoshimasa Tsuruoka[4]

Abstract: On the one hand, in the domain of imperfect information games, agents often traverse a game tree and use its transition rules to calculate a strong strategy profile in a large imperfect information games in methods developed in recent years like counterfactual regret minimization or DeepStack. On the other hand, in the domain of reinforcement learning, agents do not use the transition rules but learn from interaction to the environment to calculate a policy in a stational single-agent game.
In this paper, we propose a model for solving a multi-player imperfect information game without knowing its transition rules in advance like reinforcement learning. We also apply some existing methods to the model and compare their results. The contribution of this paper is to model a task to calculate a Nash equilibrium in a multi-player imperfect information game without knowing the rules for the first time.

[1] 東京大学大学院工学系研究科電気系工学専攻
Department of Electrical Engineering and Information Systems, Graduate School of Engineering, The University of Tokyo
[2] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation
[3] 理化学研究所 革新知能統合研究センター
RIKEN Center for Advanced Intelligence Project
[4] 東京大学大学院情報理工学系研究科電子情報学専攻
Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo
[a)] kkawamura@logos.t.u-tokyo.ac.jp
[b)] suzuki.jun@lab.ntt.co.jp

1. はじめに

人工知能分野の研究対象として，ゲーム AI が盛んに用いられている．その理由の一つは，実世界の問題にいきなり人工知能の技術を適用しても実世界の問題が複雑すぎて技術の到達点（マイルストーン）を評価するのが困難となるため，行動の制約や報酬などのルールが厳密に定められたゲームにおいて技術を確立し，徐々に複雑なゲーム設定へと対象を移行することで，技術を効果的に高めていくためである．従って，現実の問題設定により近い不完全情報

ゲームを対象とし，高い性能を発揮するプレイヤを作ることができれば，実問題への人工知能技術の応用の可能性を大きく進展できると考えられる．

不完全情報ゲームの文脈では，Texas Hold'em というポーカーゲームがよく題材として用いられる．Bowling らは，Texas Hold'em のベット額を制限した heads-up limit Texas Hold'em (HULHE) の解を，counterfactual regret minimization+ (CFR+) [1] という手法を用いて求めることに成功した (essentially weakly solved) [2]．また，Moravčík らは，相手の手を仮定した時の最終的な期待報酬をニューラルネットワークで予測して CFR+を用いる DeepStack [3] というアルゴリズムを提案し，HULHE より複雑なゲームである heads-up no-limit Texas Hold'em (HUNL) においてプロのポーカープレイヤに勝利した．ほぼ同時期に，Brown らは，抽象化したゲームを探索して相手の手に応じて部分ゲームを生成し CFR+を用いて解くことで戦略を得る，nested subgame solving [4] と呼ばれるアルゴリズムを提案し，このアルゴリズムを用いた AI である Libratus は HUNL においてプロのポーカープレイヤに勝利した [5]．

これらのアルゴリズムは，いずれもゲームの木を探索することで戦略を求める手法である．したがって，これらの手法を適用するには，agent があらかじめゲームの内容や状態遷移規則を全て知っておく必要がある．

この，事前にゲームの規則を知っておかなければならないという制約を用いない AI が成功している分野として，Atari が挙げられる．Atari 2600 は 1 人用のビデオゲームであり，Mnih らは，強化学習 [6] の一種である Q 学習 (Q-learning) [7] に深層学習を適用した deep Q-network (DQN) [8] と呼ばれる手法によって，事前にゲームのルールを知ることなく，画像のみを入力として同一のモデルで，Atari 2600 に含まれる 49 のゲームにおいて高いスコアを出すことに成功した．

このような強化学習の設定では，Markov decision process (マルコフ決定過程，MDP) で表される環境に対して，agent が環境から観測と報酬を受け取り，環境に対してアクションを行うことで環境の状態が遷移する，という流れを繰り返すことでゲームが行われる．この際，ゲームの状態遷移規則は agent に必ずしも明示的には与えられず，agent が自力で学習していく設定がより一般的である．現実の問題に適用するにあたっても，問題の内容や状態遷移規則が全て分かっているという状況は考えにくく，DQN のように未知の環境と相互作用しながら agent が状態遷移規則を把握していくような設定で問題を解くことができる手法が望まれる．

これらのことから本研究では，複数人の不完全情報ゲームにおいて未知の環境で agent が学習を行うという学習方式を提案し，定式化する．また，本論文ではこの「定式化された学習の枠組み」を指してモデルと呼ぶ．

表 1 提案モデルの立ち位置

		1 人ゲーム	2 人以上のゲーム
既知の環境	モデル例	MDP	展開型ゲーム
	適用手法例	線形計画法	CFR+
未知の環境	モデル例	強化学習	**提案モデル**
	適用手法例	Q 学習	

本研究で提案するモデルの立ち位置を表 1 に示す．

このモデルには，CFR+や DeepStack などはそのままでは適用できない．先述の通り，CFR+などを適用するためにはゲームの状態遷移規則が必要であり，学習初期の段階では agent はこの規則を知らないからである．また，DQN などの強化学習の手法もそのままでは適用することができない．なぜなら，通常の強化学習では環境を定常的なマルコフ決定過程だと仮定しているが，今回の設定では agent が複数人いて非定常な意思決定を行うため，環境を定常的であるとみなせないからである．

このモデルを解くことができるということは，中身がわからない（ゲーム木を直接構成できない）ゲームであって，ゲーム中にも一部の agent からしか観測できないような情報が出てくるとしても，事前にプレイすることができれば解くことができるということである．ゲーム AI を現実の問題に応用するにあたって，実問題の状態遷移規則は多くの場合わからないため，このモデルを解くことは重要なタスクであると言える．例えば，現実世界における会話・交渉においては，ルールや報酬関数などが明示されているわけではなく，相手によって取る戦略も変わるため，CFR などを用いて解を求めることはできない．しかしながら，人間は様々な相手と対話を繰り返すうちに，相手の情報を探りながら自分の意見を伝える方法を身に着けていく．このモデルは通常の不完全情報ゲームのモデルに比べてよりこのような状況に近い設定である．

Heinrich らは，CFR+などの regret matching を用いた手法とは別のアプローチで不完全情報ゲームに取り組み，強化学習で相手の戦略に対する最適応答戦略を近似してこれを教師あり学習で平均化することによりナッシュ均衡戦略を求める，fictitious self-play (FSP) [9] と呼ばれる手法を提案した．この手法では，ゲームの状態遷移規則などを用いることなく戦略を求めるので，本研究で取り扱う設定にもこの手法をそのまま適用することができる．しかしながらこの手法は，CFR+などの手法と比べて収束が遅く，収束先の戦略もナッシュ均衡から比較的遠くなってしまう．

また，Lanctot らは，CFR の収束性を保ちつつサンプリングを行う，Monte Carlo CFR (MCCFR) [10] と呼ばれる CFR を拡張した手法を提案した．その特殊形である outcome-sampling MCCFR では，全 agent について戦略を固定してサンプリングを行い，得られた系列についてのみ学習を行うことで，ゲームの状態遷移規則を用いること

表 2　用語の統一

意味	強化学習	不完全情報ゲーム	本論文
ゲームを行う主体	agent	player	agent
agent の意思決定に用いる確率分布	方策 (policy)	戦略 (strategy)	戦略
agent が最大化する対象	報酬 (reward)	利得 (utility)	報酬

なく CFR と同等以上の解を求めることに成功している．しかしながら，この手法は各状態に対して戦略のテーブルをもつ必要があるため，大規模なゲームには適用できないという問題がある．

本論文では，未知の多人数不完全情報ゲームにおいて agent が学習を行うというモデルを提案し，このモデルに適用可能だと考えられる上記の手法について，小規模な二人零和のゲームにおいて比較実験を行う．本研究の貢献として，多人数不完全情報ゲームについてゲームの状態遷移規則を事前に用いることなくナッシュ均衡戦略を求めるタスクを，我々の知る限り初めてモデル化したことが挙げられる．

2. 関連研究

2.1 用語の統一

本論文では，強化学習と不完全情報ゲームという 2 つの異なる文脈にまたがって議論を行うため，似た意味の異なる用語がいくつか登場する．混乱を避けるため，本論文では表 2 の通りに用語を統一して用いる．各文脈における各用語の具体的な定義は後述する．

2.2 MDP

MDP は，Markov chain (マルコフ連鎖) に agent の行動と報酬を加えたモデルである．MDP は状態集合 S，行動集合 A，状態遷移関数 T，報酬関数 R の組からなり，agent が状態 $s \in S$ において行動 $a \in A$ を取ると，状態は s と a によって $T(s,a)$ で与えられる確率分布にしたがって次状態 $s' \in S$ に遷移し，s と a によって $R(s,a)$ で与えられる分布にしたがって報酬 $r \in \mathbb{R}$ が agent に与えられる．MDP においては，状態遷移関数 T と報酬関数 R は直前の状態 s と行動 a のみに依存し，それ以前の状態や行動，時刻などには依存しない (マルコフ性を持つ)．

MDP において，agent の戦略 σ を，状態 s に対してその状態から各行動 $a \in A$ を選択する確率分布を返す関数とする．agent は以下の式で表される累積報酬を最大化することを目的とする．

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right] = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\ R(s_t,a_t)|_{a_t\sim\sigma(s_t)}\right] \qquad (1)$$

ここで，t は時刻であり，γ は割引率 (discount factor) と呼ばれる 1 以下の正の実数である．時刻 $t=0$ において環境の状態は s_0 であり，agent が戦略 σ に従って各時刻 t で行動 a_t を生成することで環境の状態が遷移していく．

MDP では状態 s と agent の意思決定に用いられる観測 ($\sigma(o,a)$ における o) が同一視されているが，環境を完全には観測できないモデルとして partially observable Markov decision process (部分観測マルコフ決定過程，POMDP) が用いられることもある．POMDP では，MDP に加えて観測確率関数 O があり，状態 s における観測 o は確率分布 $O(s,a)$ にしたがって生起する．

2.3 強化学習

強化学習とは，agent が環境と相互作用しつつ最適な戦略を学習するモデルである [6]．基本的な強化学習のモデル [11] では，環境 $\mathcal{E}$ は状態 s を持ち，s に関係する観測 o が agent に与えられる．agent は o を受けて行動 a を選択し，その行動に応じて環境 $\mathcal{E}$ は状態遷移を行い，agent に報酬 r を与える．これを繰り返し，agent は観測 o から行動 a を選択する戦略 σ のうち，累積報酬を最大化するものを学習することを目指す．

この文脈における強化学習では，agent の学習に用いる情報は agent が観測できる o_t, a_t, r_t のみである．すなわち，環境の状態遷移関数 T や報酬関数 R などを陽に知ることはできず，実際にゲームを行なった軌跡から最適な戦略を学習することになる．

強化学習の環境としては，先述の MDP あるいは POMDP が用いられることが多い．しかしながら，これらのモデルでは，状態遷移などが時刻に依存する環境を記述することができないという問題がある．

2.4 Q 学習と DQN

Q 学習 [7] は，MDP を強化学習の設定で解く手法の一つである．Q 学習では，状態 s と行動 a に対して，その状態から行動 a を行い，その後最適な行動を行なった場合の累積報酬を推定するように学習を行う．すなわち，

$$Q^*(s_t,a_t) = \mathbb{E}\left[R(s_t,a_t) + \sum_{\tau=1}^{\infty}\gamma^\tau \max_{a_{t+\tau}} R(s_{t+\tau},a_{t+\tau})\right] \qquad (2)$$

となる Q^* を学習することを目指す．最終的な戦略は，学習された $Q(s,a)$ に対して，$Q(s,a)$ を最大にする a のみ確率 1 で選択するようにする．

Mnih らは，Q 学習に深層学習の手法を適用した DQN [8] を提案した．DQN では，Q 学習における $Q(s,a)$ をニューラルネットワークを用いて近似することで，状態空間をテーブルで持つことができないような大規模なゲームに対しても Q 学習を適用できるようになっている．

2.5 多人数ゲームと展開型ゲーム

2.5.1 1人ゲームと多人数ゲーム

前項までで述べたMDPなどのモデルは，いずれもagentが1人で意思決定を行うモデルであった．

複数のagentが情報を共有しながら協力するゲームであれば，全てのagentの状態・行動空間を掛け合わせた1人のagentを操作するゲームだとみなすことでMDPに帰着させることができる [12] が，複数のagentが独立に行動する (情報を共有しない) ゲームである場合，各agentから見た環境はMDPにはならない．なぜなら，他のagentが戦略 σ を変更すると，別のagentから見た環境の状態遷移関数 T も変更されることになり，環境が時刻に対して定常ではなくなるからである．したがって，二人以上のゲームを考える場合，別のモデルが必要となる．

2.5.2 有限展開型ゲーム

有限展開型ゲーム (finite extensive-form game) は，複数人のターン制の不完全情報ゲームを記述するモデルである．有限展開型ゲームはagentの集合 N，行動の集合 A，状態に対応する履歴 (history) の集合 H，終端履歴 (terminal history) の集合 Z，履歴に対してそのターンに行動するagentを与える関数 P，チャンスノードの確率分布を与える関数 f_c，各agent $i \in N$ の報酬関数 U_i，そして情報分割 (information partition) $\mathcal{I}_i$ からなる [13].

ゲーム開始時点で，履歴 h は空列 $h = \varepsilon$ である．各履歴 $h \in H \setminus Z$ において，agent $i = P(h) \in N \cup \{c\}$ は行動 a を一つ選択する．ただし，$P(h) = c$ の時は確率分布 $f_c(h)$ にしたがって a が選択される．ゲームの履歴は $h' = ha$ に遷移する．終端履歴 $h = z \in Z$ に到達した場合，それ以上履歴の遷移は行われず，各agentに報酬 $U_i(z)$ が与えられる．

不完全情報ゲームと完全情報ゲームの違いは，不完全情報ゲームにはいずれかのagentから区別できないノードが存在するということである．このような，agent i から見て区別できないノードの集合を，情報集合 (information set) $I_i \in \mathcal{I}_i, I_i \subset H$ と呼ぶ．

2.5.3 Kuhn Poker

不完全情報ゲームの例として，ポーカーゲームの一種であるKuhn poker [14] を用いて説明する．Kuhn pokerでは，2人のagentと3枚の相異なる数字が書かれたカード (ここでは1, 2, 3とする)，各agentに数枚のチップを用意する．まず，各agentは相手から見えないようにカードを1枚ずつ引き，チップを場に1単位ずつ供託する．agent1はチップを追加する (Bet) か，様子を見る (Check) を選択する．agent1がCheckを選択した場合，agent2も同様の選択を行い，両agentがCheckを選択するとゲームは終了する．片方のagentがBetを選択した場合，そのagentは1単位のチップを追加で供託する．もう片方のagentはその賭けに乗る (Call) かゲームを降りる (Fold) を選択し，どちらを選んでもゲームは終了する．ゲームが終了すると，Foldを選択していないagentのうち，最も持っているカードが大きいagentが場の全てのチップを受け取る．例えば，各agentのカードが $(2, 3)$ であり，たどり着いた終端履歴が $(\mathrm{Check}, \mathrm{Bet}, \mathrm{Call})$ であった場合，agent1は2単位のチップを失い，agent2は2単位のチップを得る．

このゲームにおいて，例えば各agentのカードが $(2, 3)$ である場合と，$(2, 1)$ である場合を，agent1から区別することはできない．したがって，例えば履歴 $h = ((2, 1), \mathrm{Check}, \mathrm{Bet})$ と $h' = ((2, 3), \mathrm{Check}, \mathrm{Bet})$ は，agent1にとっては同じ情報集合 I_1 に属する $(h \in I_1, h' \in I_1)$ が，agent2にとっては異なる情報集合に属する $(h \in I_2, h' \in I'_2, I_2 \neq I'_2)$.

2.5.4 戦略とナッシュ均衡

展開型ゲームにおける各agentの目標は，各情報集合に対して行動の確率分布を与える戦略 $\sigma_i(I)$ を考え，以下の式で表される期待報酬を最大化するような戦略を求めることである．

$$u_i(\sigma) = \mathbb{E}\left[U_i(z)\right] = \sum_{z \in Z} U_i(z) p^{\sigma}(z) \tag{3}$$

ただし，ここで $p^{\sigma}(h)$ は各agentが戦略プロファイル σ に従って行動したときに $h \in H$ に到達する確率であり，戦略プロファイル $\sigma = \{\, \sigma_i \mid i \in N \,\}$ は各agentの戦略 σ_i を組み合わせたものである．

この状況では，期待報酬が他のagentの戦略に依存しているので，agent i についての最適な戦略は一意に定まらない．すなわち，相手の戦略 σ_{-i} を固定すれば最適な戦略

$$b_i\left(\sigma_{-i}\right) = \sigma_i^* = \arg\max_{\sigma_i} u_i\left(\sigma_i, \sigma_{-i}\right) \tag{4}$$

を求めることはできるが (この戦略 b_i を戦略 σ_{-i} に対する最適応答戦略 (best response strategy) と呼ぶ)，agentは相手の戦略を陽に観測できないので，「戦略が最適であるかどうか」は一意に定まらない．

従って，戦略の最適性を一意に定めるために，一度ゲームが2人であると仮定して，「自分に対する最適応答戦略」に対する報酬を最大化するような戦略を「最適」であるとする．すなわち，戦略 σ_i^* が「最適」であるとは，

$$\sigma_i^* = b_i\left(b_{-i}\left(\sigma_i^*\right)\right) \tag{5}$$

が成り立つことであるとする．しかし，この定義はagentが3人以上になったとき成り立たない ($-i$ が複数人を指すようになるため)．そこで，「どの agent $j \neq i \in N$ も，$-j$ に対する最適応答戦略を取っているような戦略プロファイル」に対する最適応答戦略を「最適」であるとする．すなわち，

$$\sigma_i^* = b_i(\sigma_{-i}^*),\ \sigma_{-i}^* = \{\, \sigma_j^* \mid \forall j \neq i,\ \sigma_j^* = b_j(\sigma_{-j}^*) \,\} \tag{6}$$

となるような σ_i^* を「最適」な戦略であるとする．明らか

に，式 (6) は各 agent について対称で，

$$\forall i \in N,\ \sigma_i^* = b_i(\sigma_{-i}^*) \tag{7}$$

と書き直せる．戦略プロファイル σ が上式を満たすとき，戦略プロファイル σ はナッシュ均衡 (Nash equilibrium) であると言う．不完全情報ゲームにおいては，ナッシュ均衡戦略を求めることが目的の一つとなる．また，式 (7) の条件を緩和した

$$\forall i \in N,\ \sigma_i^* + \varepsilon = b_i(\sigma_{-i}^*) \tag{8}$$

を戦略プロファイル σ が満たすとき，戦略プロファイル σ は ε-ナッシュ均衡であると言う．

戦略プロファイルがナッシュ均衡にどれだけ近いかを定量的に表す指標として，可搾取量 (exploitability) [15] がある．一般の (3 人以上のゲームも含む) 零和不完全情報ゲームにおける，戦略プロファイル σ に対する可搾取量は次式で計算できる [16].

$$\varepsilon(\sigma) = \sum_{i \in N} u_i(b_i(\sigma_{-i})) \tag{9}$$

すなわち，この値は「相手の agent の戦略を固定したとき，その相手からどれだけ搾取することができるか」を表している．可搾取量 $\varepsilon(\sigma) \geq 0$ は σ がナッシュ均衡であるとき，またそのときに限り 0 になる．また，可搾取量が ε 以下であるような戦略プロファイルは少なくとも ε-ナッシュ均衡となる．従って，可搾取量が 0 に近いほどその戦略プロファイルはナッシュ均衡に近いと言える．

2.6 CFR

Q 学習が MDP を強化学習の設定で解く手法であるように，CFR [13] は展開型ゲームをゲームの状態遷移規則などを全て用いてよい設定の元で解く手法である．CFR は regret matching と呼ばれる手法を元にしており，ゲームの木を全探索して各情報集合における counterfactual regret を最小化するように学習を行うことで，全体の regret を最小化することができ，二人零和ゲームなどにおいてナッシュ均衡戦略を求めることができる．

Lanctot らが提案した MCCFR [10] は，CFR における counterfactual regret の計算をサンプリングを用いて行えるようにした手法である．その一種である outcome-sampling MCCFR では，サンプリング戦略から終端履歴 $z \in Z$ を生成し，その終端履歴についてのみ計算を行う sampling counterfactual regret $\tilde{r}$ の期待値 $\mathbb{E}[\tilde{r}]$ が CFR における counterfactual regret r に等しくなることを示し，$\tilde{r}$ を最小化するように学習を行うことで，ゲーム木全体を探索することなくナッシュ均衡戦略を求めることができる．

また，Tammelin は，CFR を改良した CFR+ [1] と呼ばれる手法を提案した．この手法は Johanson らの用いた public chance sampling CFR [17] の regret の計算に改良を加えたもので，CFR と比べて収束の速度が優れている．

2.7 FSP

不完全情報ゲームを解く別の手法に，fictitious play (FP) [18] がある．FP では，平均戦略に対する最適応答戦略を求め，これまでに求めた最適応答戦略の平均を平均戦略とする，という操作を繰り返すことで，平均戦略が二人零和ゲームなどにおいてナッシュ均衡戦略に収束する．この手法は標準型ゲームと呼ばれる展開型ゲームとは異なるモデル上で行う手法であるため，ポーカーなどのゲームに適用するには問題があった (任意の展開型ゲームは標準型ゲームで表現することができるが，その際状態数がゲーム木の深さに対して指数的に増加するため，実用的ではなかった)．Heinrich らはこの問題を解決した full-width extensive-form FP (XFP) [9] を提案し，さらにその手法を sample-based に変更した FSP を提案した．FSP にニューラルネットワークを適用した neural fictitious self-play (NFSP) [19] は，ポーカーゲームの一種である heads up limit Texas Hold'em (HULHE) において，既存のアルゴリズムに匹敵する性能を発揮した．

FSP は，最適応答戦略の計算のためにゲーム木全体を探索しなければならないという XFP の欠点を解決し，HULHE のようなゲーム木全体をメモリに載せることが難しい大規模なゲームにおいても戦略を計算できるようにした手法である．しかしながら，最適応答戦略の計算に強化学習を，平均戦略の計算に教師あり学習を用いたことで，理論的な収束保証は失われてしまっている．

3. 提案モデル

本研究では，MDP や POMDP で表される定常な環境における強化学習に対応するモデルとして，展開型ゲームで表される多人数不完全情報ゲームにおけるモデルを提案する．

MDP における強化学習は，学習アルゴリズム Alg と環境 $\mathcal{E} = < S, A, T, R >$ を用いて次のように定式化できる．まず，環境 $\mathcal{E}$ は内部状態 $s \in S$ を保持しており，学習アルゴリズム Alg は $\mathcal{E}$ の行動集合 A のみを知っている．$\mathcal{E}$ は内部状態 s から計算される確率分布 $R(s)$ にしたがって報酬 r を生成し，Alg に s, r を伝える．Alg は $a \in A$ を $\mathcal{E}$ に送信し，$\mathcal{E}$ は確率分布 $T(s, a)$ にしたがって次状態 s' に遷移する．Alg が知ることのできる情報は，行動集合 A と系列 (s_t, r_t, a_t)，そして $\mathcal{E}$ が MDP で表せることのみである．Alg の最終的な目標は，状態 s から行動 A 上の確率分布を与える戦略 σ のうち，累積報酬を最大化するものを学習することである．このモデルの模式図を，図 1 に示す．

この設定に倣い，多人数不完全情報ゲームにおける強化学習を次のように定式化する．まず，環境 $\mathcal{E}_x = <$

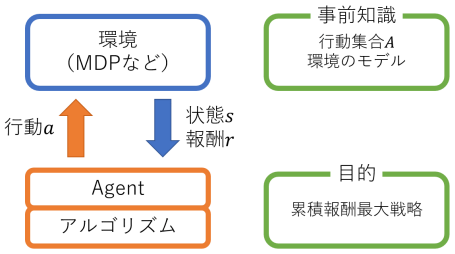

図 1 強化学習の模式図

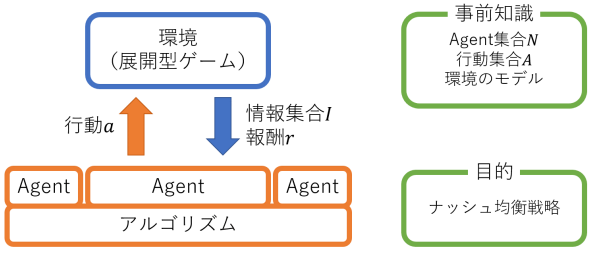

図 2 提案モデルの模式図

$N, H, A, P, f_c, \mathcal{I}_i, Z, U_i >$ は内部状態 $h \in H$ を保持しており，ゲーム開始時においては h は空列 ε である．学習アルゴリズム Alg は最初，$\mathcal{E}_x$ の行動集合 A と agent の集合 N を知っている．$\mathcal{E}_x$ は内部状態 h から行動する agent $P(h)$ を計算し，情報分割 $\mathcal{I}_{P(h)}$ と内部状態 h を用いて情報集合 $I_{P(h)}$ を計算し，Alg に $P(h)$ と $I_{P(h)}$ を伝える．$P(h) \in N$ の場合，Alg は $a \in A$ を $\mathcal{E}$ に送信し，$\mathcal{E}_x$ は次状態 $h' = ha$ に遷移する．$P(h) = c$ の場合は Alg は行動を送信せず，$\mathcal{E}_x$ は確率分布 $f_c(h)$ にしたがって行動 a を生成し，次状態 $h' = ha$ に遷移する．$h \in Z$ の場合は $P(h)$ は値を取らず，Alg に確率分布 $U_i(h)$ にしたがって報酬 $u = \{ u_i \mid \forall i \in N \}$ が送信される．Alg が知ることのできる情報は，行動集合 A，agent の集合 N，系列 $(P_t(h), I_{P_t(h),t}, a_t)$，最終的な報酬 u，そして $\mathcal{E}$ が展開型ゲームで表せることのみである．Alg の最終的な目標は，観測 $(P(h), I_{P(h)})$ から行動 A 上の確率分布を与える戦略 σ のうち，可搾取量を最小化するもの (ナッシュ均衡戦略) を学習することである．このモデルの模式図を，図 2 に示す．また，このモデルの強化学習との比較を表 3 に示す．

このモデルにおいては，Alg として CFR などを適用することはできない．なぜなら，CFR などの手法はゲームの木を再帰的に探索する必要があるためである．例えば CFR の場合，履歴 h における counterfactual value の計算には，h から行えるすべての行動 a に対して ha における counterfactual value の計算を行う必要があるが，このモデルでは環境の状態遷移は不可逆であり，Alg が環境に対して送ることのできる行動も 1 種類だけなので，すべての ha を観測することはできないからである．

強化学習に用いることのできる Q 学習などは，このモデルに直接適用することができる．しかし，このモデルでは

表 3 強化学習と提案モデルの比較

	強化学習	提案モデル
環境	MDP $< S, A, T, R >$	展開型ゲーム $< N, H, A, P, f_c, \mathcal{I}_i, Z, U_i >$
事前知識	行動集合 A 環境のモデル	agent 集合 N 行動集合 A 環境のモデル
内部状態	状態 s	履歴 h
alg に与えられる情報	状態 s 報酬 r	情報集合 $I_{P(h)}$ 行動する agent $P(h)$
alg の送信する情報	行動 a	行動 a
目的	累積報酬最大戦略	ナッシュ均衡戦略

自分以外の agent を環境の一部として考えた場合，環境が時間的に定常ではないため，正しく最適戦略を求めることはできない．また，自分以外の agent を固定して考えた場合，Q 学習は強化学習の目的である累積報酬を最大化する戦略を求めることができるが，これは相手に対する最適応答戦略に等しく，ナッシュ均衡解を求めることはできない．

また，通常の強化学習で得られる戦略の平均を取ることで FP を近似する手法である FSP は，このモデルに直接適用することができ，近似的なナッシュ均衡解を得られることがいくつかのゲームで実験的にわかっているが [9]，ナッシュ均衡解に収束することが理論的に証明されているわけではない．

outcome-sampling MCCFR は，ここで挙げた手法の中で唯一ナッシュ均衡解に収束することが証明されている手法である．この手法は終端履歴をサンプリング用の戦略プロファイルからサンプリングすることで，木探索をすることなく CFR とほぼ同等の収束速度を得ることができる．しかしながら，FSP や DQN などと異なり，関数近似をする方法が確立されていないため，メモリに載せられないような大規模なゲームには適用することができないという問題点がある．また，論文中で用いられている lazy updating によって，累積戦略の計算に多少の誤差が生じるため，理論通りの結果が出ることが保証されていないという問題点もある．

このモデルは，通常の不完全情報ゲームで用いられる，ルールを全て知っている状態から戦略を計算するモデルと比べて，より難しく現実の問題に近い設定であると考えられる．例えば，現実世界における会話・交渉においては，ルールや報酬関数などが明示されているわけではなく，相手によって取る戦略も変わるため，CFR などを用いて解を求めることはできない．しかしながら，人間は様々な相手と対話を繰り返すうちに，相手の情報を探りながら自分の意見を伝える方法を身に着けていく．このモデルは通常の不完全情報ゲームのモデルに比べてよりこのような状況に近い設定であると言える．

表 4　各ゲームの情報集合数

Kuhn poker	Leduc Hold'em	HULHE	NLHE
10^1	10^2	10^{13}	10^{160}

表 5　各ゲームにおけるランダム戦略の可搾取量

Kuhn poker	Leduc Hold'em
0.917	4.77

4. 実験

この多人数不完全情報ゲームにおける強化学習のモデルに対して，Q 学習，FSP，NFSP，および MCCFR が，どの程度ナッシュ均衡解に近い戦略プロファイルを得られるのかを確かめるために，実験を行った．実験では，不完全情報ゲームとして Kuhn poker [14] および Leduc Hold'em [20] を用いた．Kuhn poker は 2 章で述べた通り，2 人の小規模な不完全情報ゲームである．Leduc Hold'em は，public card を 1 枚追加し，Kuhn poker をより通常の Hold'em に近づけた 2 人不完全情報ゲームである．これらのゲームの複雑さの指標として，情報集合数の一覧を表 4 に示す．Kuhn poker 及び Leduc Hold'em は，Hold'em の重要な要素を残しつつ複雑性を減らしたゲームであると言える．

実験の対象としてこれらの単純化したゲームを用いたのは，可搾取量によって戦略を定量的に評価するためである．可搾取量は第 2 章で述べた通り，与えられた戦略プロファイルがどの程度ナッシュ均衡戦略から離れているかを定量的に示す正の値で，最適応答戦略に 1 ゲームあたりいくら搾取されるかを表している．可搾取量の計算にはゲームの木を探索する必要があるため，可搾取量の時間計算量はゲームの情報集合数に比例する．したがって，この値は大規模なゲームに対しては計算することができない．本論文では，可搾取量をグラフにプロットして戦略を定量的に評価するため，情報集合数の少ない単純化したゲームを用いた．

実験において，FSP の強化学習部分には Q 学習を，教師あり学習には行動の単純平均を用いた．Q 学習および FSP の強化学習部分では，学習率を 0.1，ε-greedy における ε を $\varepsilon = 0.1$ とした．これ以外の FSP および NFSP のハイパーパラメータは [16] と同じ設定にした．MCCFR のサンプリング戦略では $\varepsilon = 0.6$ とした．

Kuhn poker における実験の結果を図 3 に，Leduc Hold'em における結果を図 4 に示す．また，可搾取量の目安を示すため，各情報集合において合法手をランダムに選択する戦略プロファイルの可搾取量を表 5 に示す．

Q 学習は相手の戦略に対する最適応答戦略を求めようとするため，戦略が収束したとしてもナッシュ均衡解を求められるとは限らない．そのため，可搾取量も非常に大きい値に収束している．FSP と MCCFR は，学習が進むにつれて可搾取量が減少しており，提案モデルに適用可能であることがわかる．NFSP も学習が進むにつれて可搾取量が減少しているが，Kuhn poker では約 5×10^5 回，Leduc Hold'em では約 2×10^6 回の episode 付近で可搾取量の減少が止まっていることがわかる．これは，NFSP がニューラルネットワークによる関数近似を行っているため，関数近似を行っていない他の手法に比べて精度が落ちてしまったからではないかと考えられる．

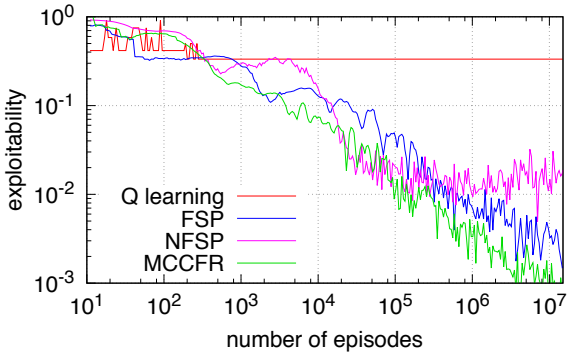

図 3　Kuhn poker における可搾取量

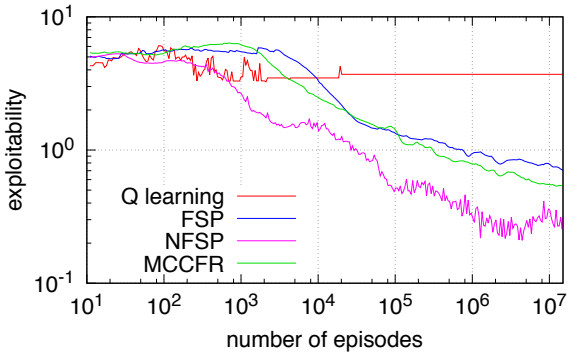

図 4　Leduc Hold'em における可搾取量

表 6　各手法の実行時間

	Kuhn poker	Leduc Hold'em
Q 学習	1.0	1.0
FSP	2.0	2.2
NFSP	5.3×10^3	4.5×10^3
MCCFR	2.0	2.0

また，この実験設定における各手法の実行時間を表 6 に示す．計測のために 10^6 episodes の実験を 10 回行い，その平均を取り，Q 学習の実行時間を 1 とする比率で示した．表 6 から分かる通り，NFSP は他の手法に比べて非常に低速であることがわかる．これは，ニューラルネットワークを用いるために学習の各ステップで大規模な行列計算が入ってしまうことが原因であると考えられる．

5. おわりに

本論文では，未知の多人数不完全情報ゲームのナッシュ均衡戦略を求めるタスクをモデル化した．また，このタス

クに適用可能な既存手法を挙げ，それらについて実際に小規模な 2 人零和不完全情報ゲームにおいて解が求められるかを検証した．その結果，FSP，MCCFR，NFSP がこのタスクに適用可能であることを実験的に示した．

今後の課題として，このモデルに対してより効率よくナッシュ均衡解を求められる手法を考案することが挙げられる．今回実験に用いた手法のうち，FSP と MCCFR は，戦略の計算を行うために各情報集合に対するテーブルをもつ必要があるため，大規模な不完全情報ゲームには適用できないという欠点がある．NFSP はニューラルネットワークによって関数近似を行うことでこの欠点を解決しているが，第 4 章で述べた通り，他の手法に比べると精度が低くなってしまう．また，FSP や MCCFR と比べるとはるかに実行時間がかかることも欠点の一つである．これらの欠点を解決するような手法を考案することが今後の課題として挙げられる．

参考文献

[1] Tammelin, O.: Solving Large Imperfect Information Games Using CFR+, *arXiv:1407.5042* (2014).

[2] Bowling, M., Burch, N., Johanson, M. and Tammelin, O.: Heads-up limit hold'em poker is solved, *Science*, Vol. 347, No. 6218, pp. 145–149 (2015).

[3] Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M. and Bowling, M.: DeepStack: Expert-level artificial intelligence in heads-up no-limit poker, *Science*, (online), DOI: 10.1126/science.aam6960 (2017).

[4] Brown, N. and Sandholm, T.: Safe and nested endgame solving for imperfect-information games, *AAAI Workshop on Computer Poker and Imperfect Information Games* (2017).

[5] Brown, N. and Sandholm, T.: Safe and Nested Subgame Solving for Imperfect-Information Games, *arXiv:1705.02955* (2017).

[6] Sutton, R. S. and Barto, A. G.: *Reinforcement learning: An introduction*, Vol. 1, No. 1, MIT press Cambridge (1998).

[7] Watkins, C. J. and Dayan, P.: Technical Note: Q-Learning, *Machine Learning*, Vol. 8, No. 3, pp. 279–292 (1992).

[8] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, pp. 529–533 (2015).

[9] Heinrich, J., Lanctot, M. and Silver, D.: Fictitious Self-Play in Extensive-Form Games, *Proceedings of ICML*, JMLR Workshop and Conference Proceedings, pp. 805–813 (2015).

[10] Lanctot, M., Waugh, K., Zinkevich, M. and Bowling, M.: Monte Carlo Sampling for Regret Minimization in Extensive Games, *Advances in NIPS 22*, pp. 1078–1086 (2009).

[11] Kaelbling, L. P., Littman, M. L. and Moore, A. W.: Reinforcement learning: A survey, *Journal of artificial intelligence research*, Vol. 4, pp. 237–285 (1996).

[12] Claus, C. and Boutilier, C.: The dynamics of reinforcement learning in cooperative multiagent systems, *AAAI/IAAI*, Vol. 1998, pp. 746–752 (1998).

[13] Zinkevich, M., Johanson, M., Bowling, M. and Piccione, C.: Regret Minimization in Games with Incomplete Information, *Advances in NIPS 20*, Curran Associates, Inc., pp. 1729–1736 (2008).

[14] Kuhn, H. W.: A simplified two-person poker, *Contributions to the Theory of Games*, Vol. 1, pp. 97–103 (1950).

[15] Johanson, M., Waugh, K., Bowling, M. and Zinkevich, M.: Accelerating Best Response Calculation in Large Extensive Games, *Proceedings of the 22nd IJCAI - Volume 1*, pp. 258–265 (2011).

[16] Kawamura, K., Mizukami, N. and Tsuruoka, Y.: Neural Fictitious Self-Play in Imperfect Information Games with Many Players, *Computer Games Workshop at IJCAI-17* (2017).

[17] Johanson, M., Bard, N., Lanctot, M., Gibson, R. and Bowling, M.: Efficient Nash Equilibrium Approximation Through Monte Carlo Counterfactual Regret Minimization, *Proceedings of the 11th AAMAS - Volume 2*, pp. 837–846 (2012).

[18] Brown, G. W.: Iterative solution of games by fictitious play, *Activity analysis of production and allocation*, Vol. 13, No. 1, pp. 374–376 (1951).

[19] Heinrich, J. and Silver, D.: Deep Reinforcement Learning from Self-Play in Imperfect-Information Games, *arXiv:1603.01121* (2016).

[20] Southey, F., Bowling, M., Larson, B., Piccione, C., Burch, N., Billings, D. and Rayner, C.: Bayes' Bluff: Opponent Modelling in Poker, *Proceedings of the Twenty-First Conference on UAI*, UAI'05, Arlington, Virginia, United States, AUAI Press, pp. 550–558 (2005).