多人数不完全情報ゲームにおける仮想自己対戦を用いた強化学習

河村 圭悟[1,a)]　水上 直紀[2]　鶴岡 慶雅[2]

概要：不完全情報ゲームにおいて，ナッシュ均衡戦略は非常に重要なテーマである．特に多人数不完全情報ゲームにおいては，ナッシュ均衡解を一般に求める方法はまだ確立されていないことから，多くの関心を集めている．2 人テキサス・ホールデムは CFR+ (Tamelin, 2014) によって解かれた (generally weakly solved) が，CFR+は空間計算量の観点から 3 人以上のテキサス・ホールデムに適用するには問題がある．本研究では NFSP (Heinrich and Silver, 2016) と呼ばれる手法を用いて，CFR+では解くことが難しい多人数不完全情報ゲームのナッシュ均衡解を求めることを目指す．本研究では，学習部分にソフトマックス回帰を用いた Fictitious Self-Play (FSP) を使用して，テキサス・ホールデムのトイゲームである 2 人クン・ポーカーにおいて FSP が近似的なナッシュ均衡解を求められることを示した．また，多人数ゲームである 3 人クン・ポーカーにおいても，FSP が近似的なナッシュ均衡解を求められることを示し，CFR+の戦略に対する FSP の戦略の平均被搾取量が減少することを示した．

Neural Fictitious Self-Play in Multiplayer Imperfect Information Games

Keigo Kawamura[1,a)]　Naoki Mizukami[2]　Yoshimasa Tsuruoka[2]

Abstract: Computing Nash equilibrium solutions is an important problem in the domain of imperfect information games. Attempts of solving the problem draw considerable attention especially in the domain of multiplayer games because there is currently no method that can calculate approximate Nash equilibrium solutions in a general setting. CFR+ (Tamelin, 2014) can be used to (essentially weakly) solve two-player limit Texas Hold'em, but it cannot be applied to large multiplayer games due to the problem of space complexity. In this paper, we use Neural Fictitious Self-Play (Heinrich and Silver, 2016) to calculate approximate Nash equilibrium solutions for multiplayer imperfect information games that CFR+ can hardly solve. We show that Fictitious Self-Play (FSP) with a softmax regression allows us to calculate approximate Nash equilibrium solutions in two-player Kuhn poker and three-player Kuhn poker. We also show that the exploitability of the FSP solution by that of CFR+ decreases.

1. はじめに

人工知能分野の研究対象としてゲーム AI が盛んな理由の一つには，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高い性能を発揮する人工知能を開発することで，より不確定的で制限の少ない現実世界の諸問題に応用できるようにするためということが挙げられる [1]．従って，現実の問題設定により近いと考えられる多人数不完全情報ゲームにおいて，高い性能を発揮するプレイヤを作ることが重要である．

従来の多人数不完全情報ゲーム AI の多くは，そのゲームに熟練した人間がモデルやヒューリスティックスなどを考案することで実現されてきた [2]．しかし，現実の問題を解決するにあたって，人間が問題の抽象化などを行うのはコストが高く，コンピュータがデータや制約条件から自動的に最適戦略を導けることが望ましい．また，問題の解

[1] 東京大学工学部電子情報工学科
Department of Information and Communication Engineering, The University of Tokyo
[2] 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
[a)] kkawamura@logos.t.u-tokyo.ac.jp

決にあたって常に教師データが存在するとは限らず，そのようなデータを収集するコストも考慮すると，強化学習によってコンピュータが前提知識なしに最適戦略を求められるようになることが重要な目標である．

Tammelin らは，Counterfactual Regret Minimization+ (CFR+) と呼ばれる手法を用いて，代表的な不完全情報ゲームである 2 人リミットテキサス・ホールデム (heads-Up limit texas hold'em) のナッシュ均衡解を，抽象化などを行わずに求めることに成功した [3]．この手法は Lanctot らが提案した CFR [4] に改良を加えたものであり，不完全情報ゲームのゲーム木を完全探索するものである．しかし，この手法は時間・空間計算量が大きいため，3 人以上のテキサス・ホールデムについては，少なくとも現時点ではナッシュ均衡解を求めることができない．

Heinrich らは，不完全情報ゲームで古くから用いられてきた仮想プレイ (Fictitious Play, FP) と呼ばれる手法を応用した，仮想自己対戦 (Fictitious Self-Play, FSP) という手法を提案した [5]．この手法は，従来標準型ゲーム (Normal Form Game) で表現されたゲームにしか用いることができなかった FP を，展開型ゲーム (Extensive Form Game) で表現されたゲームにも適用できるようにしたものである．さらに，強化学習と教師あり学習を用いて適切な近似を行うことで，具体的な機械学習の方法に依存しない一般的な手法でありながら，FP と同じ収束保証性を保つことに成功している．この FSP に，具体的な学習方法であるニューラルネットワークを用いた教師あり学習と，ニューラルネットワークを Q 学習に応用した Deep Q-Network (DQN) [6] を用いた，Neural Fictitious Self-Play (NFSP) は，2 人リミットテキサス・ホールデムにおいて事前の抽象化や簡略化といった前提知識を用いることなく既存のプレイヤに匹敵する性能を発揮した [2]．

本研究の目的は，CFR+では取り扱いが難しい多人数不完全情報ゲームについて，NFSP を用いることで前提知識なしに近似的なナッシュ均衡解を求めることである．CFR+はゲーム木を全て展開し完全探索を行うため，空間計算量の観点から情報集合数が大きすぎるゲームへの適用は難しい．一方，NFSP はゲームの状態を抽象化して学習を行うため，ゲーム木の全状態を記憶する必要がなく，情報集合数が十分大きなゲームに対しても適用することができる．従って，本研究によって，従来では人間の手で情報集合数を削減しなければ取り扱うことができなかったゲームについても，近似的なナッシュ均衡解を得られるようになることが期待される．

本研究では，NFSP が多人数不完全情報ゲームのナッシュ均衡解を求められることを示すため，トイゲームであるクン・ポーカーについて NFSP・CFR+を適用して比較し，NFSP が 3 人クン・ポーカーにおいて近似的なナッシュ均衡解を求められることを示した．

2. 関連研究

2.1 展開型ゲーム

有限展開型ゲーム (finite extensive-form game) [4], [7] とは，以下の要素からなるゲームである．

- プレイヤの有限集合 N．
- 履歴 (history) h の有限集合 H．H の部分集合 $Z = \{ z \in H \mid \forall h \in H, z \not\sqsubset h \}$ の要素は終端履歴 (terminal history) と呼ばれる．また，集合 $A(h) = \{ a \mid (h, a) \in H \}$ は非終端履歴 $h \in H \setminus Z$ に対して取ることの出来るアクションの集合を表す．
- ターンプレイヤを表す関数 P．P の定義域は $H \setminus Z$ であり，値域は $N \cup \{c\}$ である．$P(h)$ は履歴 h の後にどのプレイヤがアクションを行うかを表す関数であり，$P(h) = c$ のときは次のアクションはある確率分布によって決まる．
- $P(h) = c$ のときの確率分布を表す関数 f_c．f_c の定義域は $C = \{ h \in H \mid P(h) = c \}$ であり，その値はアクションの確率分布を表す確率測度 $f_c(a|h)$ である．この確率測度は異なる h に対して独立である．
- 各プレイヤ $i \in N$ に対し定まる情報分割 (information partition) $\mathcal{I}_i$．情報分割の要素である情報集合 (information set) $I_i \in \mathcal{I}_i$ は $P(h) = i$ を満たす履歴 $h \in H$ からなり，任意の異なる履歴 $h, h' \in H$ はプレイヤ i が h と h' を区別できないとき，またそのときに限り同じ情報集合に属する．$h \in I_i$ に対する $A(h)$，$P(h)$ を単に $A(I_i)$，$P(I_i)$ と書くこともある．全プレイヤが過去に自分が取ったアクションとその時の情報集合を記憶しているとき，このゲームは完全記憶 (perfect recall) ゲームであると言う．
- 各プレイヤの利得関数 (utility function) u_i．u_i の定義域は終端履歴 Z であり，値域は実数 $\mathbf{R}$ である．特に，$\sum_i u_i = 0$ であるとき，このゲームは零和展開型ゲーム (zero-sum extensive game) であると言う．

テキサス・ホールデムなど，多くの不完全情報ゲームは展開型ゲームで表現することができる．特に，2 人テキサス・ホールデムは完全記憶零和有限展開型ゲームとして表現できる．

2.2 戦略とナッシュ均衡

プレイヤ i の戦略 (strategy) σ_i とは，任意の情報集合 I_i に対して取りうるアクション $A(I_i)$ 上の確率分布を与える関数である．プレイヤ i の戦略 σ_i 全体からなる集合 Σ_i を戦略集合 (strategy set) と言う．また，各プレイヤの戦略の集合 $\{ \sigma_i \mid i \in N \}$ を戦略プロファイル (strategy profile)，あるいは単に戦略と言い，戦略プロファイルからプレイヤ i の戦略だけを除いたものを σ_{-i} と表記する．

いま，各プレイヤが戦略 σ に従って行動したときに，履

歴 h にたどり着く確率を $p^\sigma(h)$ とする．履歴 h は各プレイヤのアクションの列なので，この確率は

$$p^\sigma(h) = \prod_{i \in N \cup \{c\}} p_i^\sigma(h) = p_i^\sigma(h)\, p_{-i}^\sigma(h), \tag{1}$$

と分解することができる．ここで，$p_i^\sigma(h)$ はプレイヤ i が戦略 σ_i に従って行動するときに，$P(h') = i, h' \sqsubset h$ を満たす全ての履歴 h' について，$(h', a) \sqsubseteq h$ となるアクション a を選択する確率の積である．また，$p_{-i}^\sigma(h)$ は i 以外の全てのプレイヤ $k \neq i$ について $p_k^\sigma(h)$ を掛けあわせたものである．任意の履歴の集合 $I \subseteq H$ について，この集合に含まれる履歴のいずれかにたどり着く確率 $\sum_{h \in I} p^\sigma(h)$ を考え，これを $p^\sigma(I)$ とする．同様にして，$p_i^\sigma(I)$ および $p_{-i}^\sigma(I)$ を定義する．

プレイヤ i に対する戦略プロファイル σ の価値は，終端履歴における利得の期待値 $u_i(\sigma) = \sum_{h \in Z} u_i(h) p^\sigma(h)$ で表される．戦略プロファイル σ に対し，i 以外のプレイヤが σ_{-i} に従ったときの i の利得を最大化する戦略 σ_i' を最適応答戦略 (best response) と言い，その価値 $b_i(\sigma_{-i}) = u_i(\sigma_i', \sigma_{-i}) = \max_{\sigma_i} u_i(\sigma_i, \sigma_{-i})$ を最適応答価値 (best response value) と言う．戦略プロファイル σ が

$$\forall i \in N, \quad u_i(\sigma) \geq b_i(\sigma_{-i}), \tag{2}$$

を満たすとき，σ はナッシュ均衡 (Nash Equilibrium) であると言う．また，σ が

$$\forall i \in N, \quad u_i(\sigma) + \varepsilon \geq b_i(\sigma_{-i}), \tag{3}$$

を満たすとき，σ は ε-ナッシュ均衡であると言う．

ゲームが 2 人零和であるとき，戦略 σ がどの程度ナッシュ均衡に近いかを表す値として可搾取量 (exploitability) $\varepsilon_\sigma = b_1(\sigma_2) + b_2(\sigma_1)$ を定義する．この値が小さいほど戦略 σ はナッシュ均衡解に近いと考えられる．実際，2 人零和対称ゲームにおいては，可搾取量が ε である戦略は少なくとも ε-ナッシュ均衡であることが知られている．

また，戦略 σ_1 に対し，戦略 σ_2 が平均して 1 ゲームあたりいくら搾取されるかを表す値も exploitability と呼ばれるが，本稿ではこの値を σ_2 の σ_1 に対する平均被搾取量と呼ぶ．

2.3 CounterFactual Regret minimization

Lanctot らが提案した CFR [4] は，regret を最小化することでナッシュ均衡を求める手法である．時刻 T における regret R_i^T は

$$R_i^T = \frac{1}{T} \max_{\sigma_i^* \in \Sigma_i} \sum_{t=1}^{T} \left(u_i(\sigma_i^*, \sigma_{-i}^t) - u_i(\sigma^t) \right), \tag{4}$$

で定義される．すなわち，この値は相手の戦略を固定したときのプレイヤ i の平均利得の最大値である．いま，2 人零和ゲームにおいて regret が $R_i^T \leq \varepsilon$ を満たすとき，平均戦略 (average strategy) は 2ε-ナッシュ均衡であることが知られている．従って，regret を最小化することでナッシュ均衡を求めることができる．

情報集合 I について，先頭部分が I に含まれるすべての終端履歴の集合を $Z_I = \{ z \in Z \mid \exists h \sqsubseteq z, h \in I \}$ とし，その要素 $z \in Z_I$ について I に含まれる先頭部分を $z[I]$ とする (このような $h = z[I] \in I$ は完全記憶ゲームにおいては高々 1 つしかない)．counterfactual value $v_i(\sigma, I)$ を式

$$v_i(\sigma, I) = \sum_{z \in Z_I} p_{-i}^\sigma(z[I])\, p^\sigma(z \mid z[I])\, u_i(z), \tag{5}$$

で定義し，immediate counteractual regret $R_{i,imm}^T(I, a)$ を式

$$R_{i,imm}^T(I, a) = \frac{1}{T} \sum_{t=1}^{T} \Big(v_i(\sigma_{(I \to a)}^t, I) - v_i(\sigma^t, I) \Big), \tag{6}$$

で定義する．ただし，$\sigma_{(I \to a)}$ は，情報集合 I についてプレイヤ $P(I)$ が履歴 $h \in I$ においてはアクション a を選択し，それ以外の履歴・プレイヤは戦略 σ に従って行動するような戦略プロファイルを表す．このとき，regret と counterfacual regret の間に式

$$R_i^T \leq \sum_{I \in \mathcal{I}_i} R_{i,imm}^{T,+}(I), \tag{7}$$

が成立する (x^+ は $\max(x, 0)$ を表す)．従って，ナッシュ均衡である戦略を求めるには，単に各情報集合に対して counterfactual regret $R_{i,imm}^{T,+}(I)$ を最小化するような戦略を計算すればよい．

以上が CFR の理論的背景である．この手法を提案した Lanctot らはチャンスノード，すなわち $P(I) = c$ なる情報集合の遷移を乱数によって実装したが，Johanson らはこれを public chance と private chance に分類し，それぞれをベクトルで表現することで性能を向上させることに成功した [8]．その後，Tammelin らが counterfactual regret を改良し提案した CFR+ [3] は，情報集合数が 10^{17} 程度ある 2 人リミットテキサス・ホールデムにおいて十分小さな ε に対して ε-ナッシュ均衡解を求めることに成功した (essentially weakly solved) [9]．

2.4 標準型ゲーム

有限標準型ゲーム (finite normal-form game) [5] とは，以下の要素からなるゲームである．

- プレイヤの有限集合 N．
- 各プレイヤ $i \in N$ の純粋戦略 (pure strategy) s_i^N の有限集合 S_i^N (展開型ゲームと区別するため，添字 N を付けている)．直積 $S^N = \prod_i S_i^N$ を戦略空間と言う．
- 各プレイヤ $i \in N$ の利得関数 u_i^N．u_i^N の定義域は S^N であり，値域は実数 $\mathbf{R}$ である．

各プレイヤが順に手番を行う展開形ゲームと異なり，全プレイヤが 1 度だけ同時に戦略を決め，その結果によって利得を得るのが標準型ゲームである．しかし，任意の展開形ゲームは事前に (展開形ゲームにおける) 戦略 σ_i を決め，それを純粋戦略 s_i^N とすることで標準型ゲームに書き換え可能であり，逆に任意の標準型ゲームは展開形ゲームに書き換え可能であるから，これらが表現するゲームの集合は等価である．

標準型ゲームにおいて, 純粋戦略に対してその純粋戦略を選ぶ確率を表す関数 $\sigma_i^N\left(s_i^N\right)$ を混合戦略 (mixed strategy) と言う．すなわち, 混合戦略は $\forall s_i^N \in S_i^N, 0 \leq \sigma_i^N\left(s_i^N\right) \leq 1$ かつ $\sum_{s_i^N} \sigma_i^N\left(s_i\right) = 1$ を満たす．混合戦略 σ_i^N 全体の集合を Σ_i^N で表す．

2.5 仮想プレイ

Brown が提案した仮想プレイ (FP) [10] や，これを一般化した Generalized Weakened Fictitious Play (GWFP) [11] は，標準型ゲームとして記述される多人数不完全情報ゲームのナッシュ均衡解を求める手法である．

ステップ t における混合戦略を $\sigma^{N,t}$ とする．いま，ステップ 1〜T までの混合戦略の平均 $\frac{1}{T}\sum_{t=1}^{T}\sigma^{N,t}$ を考え，これを $\pi^{N,T}$ とする．また，この混合戦略に対する最適応答戦略を $\beta_i^N\left(\pi_{-i}^{N,T}\right) = \max_{\sigma_i^N} u_i^N\left(\sigma_i^N, \pi_{-i}^{N,T}\right)$ とする．この戦略をステップ $T+1$ における戦略 $\sigma_i^{N,T+1}$ とすると，平均戦略 $\pi_i^{N,T}$ は

$$\pi_i^{N,T+1} = \frac{1}{T+1}\beta_i^N\left(\pi_{-i}^{N,T}\right) + \left(1 - \frac{1}{T+1}\right)\pi_i^{N,T}, \tag{8}$$

と表すことが出来る．この更新式に従って平均戦略 $\pi^{N,T}$ を更新していくのが FP である．また，GWFP は

$$\pi_i^{N,T+1} = \alpha^{N,T}\left(\beta_{i,\varepsilon^T}^N\left(\pi_{-i}^{N,T}\right) + M_T^N\right) + \left(1 - \alpha^{N,T}\right)\pi_i^{N,T}, \tag{9}$$

という更新式に従って戦略を更新する．ここで，$\beta_{i,\varepsilon}^N$ は ε-最適応答戦略を表す．$\alpha^{N,T}$，ε^T はともに $T \to \infty$ で 0 に収束し，$\sum_t^\infty \alpha^{N,t} = \infty$ を満たす数列である．M_T^N は摂動項である．式 (9) において，$\alpha_T^N = 1/T$，$\varepsilon_T = M_T^N = 0$ とすればこれは FP になる．

GWFP によって得られる平均戦略 σ_T^N は，2 人ゲームやポテンシャルゲームなど，いくつかのゲームにおいてナッシュ均衡解に収束することが証明されている．

2.6 Q 学習

Q 学習 [12] は，エージェントが環境に観測・干渉しながら報酬を最大化していく強化学習の手法である．Q 学習が扱う環境はマルコフ決定過程 (Markov decision process, MDP) で表現される．

MDP とは，以下の要素からなるモデルである．

- 状態 s の有限集合 S.
- 状態 $s \in S$ で取ることが出来るアクションの集合 A_s.
- 遷移関数 $P(s, a, s')$．$P(s, a, s')$ は状態 $s \in S$ においてアクション $a \in A_s$ を取ったとき状態 $s' \in S$ に遷移する確率である．
- 報酬関数 $R(s, a)$．$R(s, a)$ は状態 $s \in S$ においてアクション $a \in A_s$ を取ったとき環境から与えられる報酬である．

Q 学習は, MDP に対して次式で定義される行動価値関数

$$Q^*(s, a) = \max_\pi \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots\right], \tag{10}$$

を予測するように学習を行う．ここで，π は状態 $s \in S$ に対してアクション A_s 上の確率分布を与える関数であり，方策 (policy) と言う．$\mathbb{E}[\cdot]$ は期待値であり，r_t はステップ t における報酬である．$0 < \gamma < 1$ は割引率 (discount factor) であり，行動価値関数が発散するのを防ぐために導入されるパラメータである．

いま，状態 s とアクション a について関数 $Q(s, a)$ を考え，次のような更新式に従って更新する．

$$\begin{aligned} Q(s_t, a_t) \leftarrow & Q(s_t, a_t) \\ & + \alpha_t\left(R(s_t, a_t) + \gamma \max_{a \in A_{s_{t+1}}} Q(s_{t+1}, a) - Q(s_t, a_t)\right). \end{aligned} \tag{11}$$

すなわち，ステップ t で得られる報酬 $R(s_t, a_t)$ と，ステップ $t+1$ 以降で得られると考えられる報酬の最大値 $\max_a Q(s_{t+1}, a)$ を足した値に，$Q(s_t, a_t)$ が近づくように学習を行う．ここで，α_t は学習率 (learning rate) である．

Q 学習は, 学習率 α_t が $\sum_t \alpha_t = \infty$ かつ $\sum_t \alpha_t^2 < \infty$ を満たすとき，Q 値 (Q-value) $Q(s, a)$ が行動価値関数 $Q^*(s, a)$ に収束することが証明されている．

Mnih らは，この Q 学習にニューラルネットワークの技術を応用した DQN と呼ばれる手法を提案した [6]．この手法では，$Q(s_t, a_t) = Q(s_t, a_t; \theta_t)$ とパラメータ θ_t を用いて Q 値を抽象化し，次の誤差関数を最小化するように学習を行う．

$$\mathbb{E}\left[\left(R(s_t, a_t) + \gamma \max_{a \in A_{s_{t+1}}} Q(s_{t+1}, a; \theta_t) - Q(s_t, a_t; \theta)\right)^2\right]. \tag{12}$$

2.7 Extensive-form Fictitious Play

FP は標準型ゲームにおけるナッシュ均衡解を求める手法であった．任意の展開型ゲームは標準型ゲームに変形することができるが，ゲーム木の深さに対して指数的に戦略空間が広がっていくので，FP をそのまま適用するのは現実的ではない．

Heinrich らは，実現確率関数 (realization plan) と 実現等価性 (realization-equivalent) [13] という性質を用いて，展開型ゲームに対してゲーム木の深さについて線形な計算量で FP を行う extensive-form fictitious play (XFP) と呼ばれる手法を提案した [5].

情報集合 $I_i \in \mathcal{I}_i$ に対し，I_i にたどり着くまでにプレイヤ i が取るアクションの列 (sequence) を h_I^i とする．すなわち，h_I^i はプレイヤ i のアクションのみからなる列である．このような列は，ゲームが完全記憶ゲームであるときただ一つ存在する．いま，実現確率関数 $x_\sigma\left(h_I^i\right)$ を次式で定義する．

$$x_\sigma\left(h_I^i\right) = \prod_{\left(h_{I'}^i, a\right) \sqsubseteq h_I^i} \sigma_i\left(I'\right)(a). \tag{13}$$

また，この展開型ゲームを標準型ゲームに変換したときの実現確率関数を考えると，混合戦略 σ_i^N は純粋戦略の線形結合であり，純粋戦略の実現確率関数はある 1 つについてのみ 1 で他は 0 になるので，結局実現確率関数は混合戦略の線形結合の係数と等しくなる．このようにして，標準型ゲームと展開型ゲームの両方で実現確率関数という同じ概念を定義することができる．

任意の戦略 σ_{-i} と任意の履歴 $h \in H$ について，h にたどり着く確率が $p^{\sigma_i,\sigma_{-i}}(h) = p^{\sigma'_i,\sigma_{-i}}(h)$ となるとき，2 つの戦略 σ_i と σ'_i は実現等価であると言う．2 つの戦略が実現等価であることと，2 つの戦略の実現確率関数が等しいことは同値であることが知られている．また，展開型ゲームにおける任意の戦略には，その戦略と実現等価な，そのゲームを標準型ゲームに変換したときの混合戦略が存在し，この逆も成り立つ．

これらのことから，次の定理が成り立つ．2 つの戦略 π, β それぞれに対して実現等価な混合戦略 π^N, β^N と，$\lambda_1, \lambda_2 \leq 0, \lambda_1 + \lambda_2 = 1$ を満たす実数の組 λ_1, λ_2 について，

$$\mu(I) = \pi(I) + \frac{\lambda_2 x_\beta\left(h_I\right)}{\lambda_1 x_\pi\left(h_I\right) + \lambda_2 x_\beta\left(h_I\right)}\left(\beta(I) - \pi(I)\right), \tag{14}$$

で表される戦略 μ は，混合戦略 $M = \lambda_1 \pi^N + \lambda_2 \beta^N$ と実現等価である．

この定理によって，標準型ゲームにおける混合戦略の加法を展開型ゲームにおける戦略の加法で表すことができる．すなわち，第 2.5 節で述べた GWFP は

$$\pi_i^{t+1}(I) = \pi_i^t(I) + \frac{\alpha^t x_{\beta_{\varepsilon^t,i}}\left(h_I\right)}{\left(1-\alpha^t\right) x_{\pi_i^t}\left(h_I\right) + \alpha^t x_{\beta_{\varepsilon^t,i}}\left(h_I\right)}\left(\beta_{\varepsilon^t,i}(I) - \pi_i^t(I)\right), \tag{15}$$

と表現できる．この計算はゲーム木の深さに対して高々多項式程度の計算量なので，展開型ゲームで表されるゲームについても効率よくナッシュ均衡解を求めることができる．これが XFP の考え方である．

さらに，この XFP における平均戦略の計算を教師あり学習 (supervised learning) で，最適応答戦略の計算を強化学習で近似することにより，収束性を保ちながら既存の機械学習の手法を応用できるようにした手法が FSP である．また，FSP の強化学習に DQN を用い，教師あり学習にニューラルネットワークを用いて，前提知識なしに不完全情報ゲームの近似的なナッシュ均衡解を求めることに成功したものが NFSP [2] である．

3. 提案手法

3.1 提案手法概要

本研究では，2 人不完全情報ゲームのナッシュ均衡解を求める手法である NFSP を，3 人以上の多人数不完全情報ゲームに適用することを提案する．

NFSP が収束することが証明されているのは 2 人ゲーム，あるいはポテンシャルゲームのみであり，3 人以上のリミットテキサス・ホールデムでは収束する保証はない．しかし，同じ収束保証を持つ CFR は 3 人テキサス・ホールデムでも実験的に収束することがわかっており [14]，同様にして NFSP も収束することが期待される．

また，CFR+はゲームの情報集合を抽象化することなく扱うため，全ての情報集合に対する戦略を保持しなければならない．すなわち，空間計算量は情報集合数に比例する．このことから，多人数ゲーム，特にテキサス・ホールデムのような情報集合数の比較的多い多人数ゲームにおいては，CFR+は適用が難しいという問題がある．アルゴリズムによって人数を減らすことで精度を保ちながら計算を単純化する手法も提案されているが [15]，ゲームの前提知識を使わずに解くという趣旨からは外れてしまう．

従って，抽象化によって空間計算量を小さくできる NFSP を用いて，多人数ゲームのナッシュ均衡解を求めることが重要である．

3.2 NFSP による多人数ゲームの戦略計算

NFSP は，第 2.7 節で述べたとおり，XFP における平均戦略の計算をニューラルネットワークによる教師あり学習で，最適応答戦略の計算を DQN による強化学習で近似した手法である．

NFSP はゲームのプレイヤの数だけエージェントを持ち，それぞれのエージェントはゲームを行いながら学習データ $M_i = (s, a, r, s')$ をメモリ $\mathcal{M}_i$ に保存し，そのデータを用いて学習する．各エージェントは相手の平均戦略 π_{-i} に対する最適応答戦略 β_i を計算し，過去の最適応答戦略を平均した戦略に従って行動することになる．しかし，ここには各エージェントが平均戦略 π と最適応答戦略 β の両方を同時に行う必要があるという矛盾が存在する．こ

れを解決するために Heinrich らは，Shamma らが提案した dynamic fictitious play [16] の考え方を利用して，1 ステップ先の平均戦略を予測する手法を NFSP に取り入れた．すなわち，微小時刻先の平均戦略 $\pi^{t+\delta} = \pi^t + \delta \frac{d}{dt}\pi^t$ と，$\Delta\pi^t = \pi^{t+1} - \pi^t \propto \beta^t - \pi^t$ より

$$\pi^{t+1} \approx \pi^t + \eta\left(\beta^t - \pi^t\right) = \eta\beta^t + (1-\eta)\,\pi^t, \qquad (16)$$

と近似し，各エージェントはこの戦略に従ってアクションを行う．この η を anticipatory parameter と呼ぶ．

学習データの保持には，平均戦略メモリ $\mathcal{M}_i^{RL}$ に FIFO のものを用い，最適応答戦略メモリ $\mathcal{M}_i^{SL}$ に reservoir sampling [17] を用いる．また，Q 学習の戦略は ε-greedy に従って返すようにし，学習ターゲット $\theta^{Q_i'}$ は一定周期で更新するようにする．

以上をまとめた NFSP の擬似コードをアルゴリズム 1 に示す．ただし，$\Pi\left(s,a|\theta^{\Pi}\right)$ は平均戦略を計算するネットワークを，$Q_i\left(s,a|\theta^{Q_i}\right)$ は最適応答戦略を計算するネットワークを表す．また，アルゴリズム 1 は NFSP における各エージェントの相互作用と報酬の設定方法を明示するため，Heinrich らが示した fitted Q-learning を用いたアルゴリズム [2] と異なっている．

アルゴリズム 1 Newral Fictitious Self-Play (NFSP)

Require:
　ゲーム Γ，プレイヤ人数 $N \geq 2$
Ensure:
　$\Pi\left(s,a|\theta^{\Pi}\right)$ は近似的なナッシュ均衡戦略を返す

```
for i = 1, 2, ···, N do                              ▷ 各プレイヤについて
    Initialize Π_i, Q_i, M_i^SL, M_i^RL, θ^{Q'_i}, M_i
end for
for iteration = 1, 2, ··· do
    ε ← ε(iteration)                                  ▷ ε-greedy に用いる ε
    for i = 1, 2, ···, N do
        σ_i ← ε-greedy(Q_i) (確率 η) or Π_i (確率 1 − η)
    end for
    Initialize ゲーム Γ                               ▷ ゲームを初期状態に戻す
    repeat                                            ▷ ゲームが終端するまで続ける
        n ← turn player of Γ
        M_n.s ← M_n.s'
        observe 状態 s* and M_n.s' ← s*               ▷ 観測・更新
        Store M_n in M_n^RL                           ▷ 学習データを保存
        if σ_n = ε-greedy(Q_i) then
            Store M_n in M_n^SL
        end if
        Periodically update θ^{Q_n} with M ∼ M_n^RL , θ^{Q'_n}
        Periodically update θ^{Π_n} with M ∼ M_n^SL
        Periodically update θ^{Q'_n} ← θ^{Q_n}
        Sample アクション a from 戦略 σ_i
        Execute アクション a on ゲーム Γ
        M_n.r ← 0                                     ▷ 終端状態以外では報酬は 0
    until Γ is over                                   ▷ ゲームが終端するまでループ
    for i = 1, 2, ···, N do
        set 報酬 M_i.r
    end for
end for
```

4. 実験

4.1 実験概要

テキサス・ホールデムのトイゲームであるクン・ポーカー (Kuhn poker) について，FSP が近似的な最適戦略を求められることを示す．はじめに，2 人クン・ポーカーにおいて，FSP が近似的なナッシュ均衡解を求められることを示す．次に，多人数ゲームである 3 人クン・ポーカーにおいて，FSP が近似的なナッシュ均衡解を求められることを示し，イテレーションの増加に従って CFR+の戦略に対する FSP の戦略の平均被搾取量が減少することを示す．

クン・ポーカーは，テキサス・ホールデムのルールを単純化したゲームである．各プレイヤは (プレイヤの人数) +1 枚の相異なるカードの中からそれぞれ 1 枚のカードを受け取り，1 単位のチップを掛け，1 回のベットラウンドを行う．ベットラウンドが全て終了した時点でフォルドしていないプレイヤが 2 人以上居た場合，互いの手札を公開し (ショーダウン)，よりカードの数字が大きいプレイヤが全ての掛け金を回収する．1 人を除いて全員がフォルドしていた場合，残ったプレイヤが全ての掛け金を回収する．

なお，テキサス・ホールデムは，カードの枚数が 13 種類 ×4 枚=52 枚であること，各プレイヤの受け取るカードが 2 枚であること，ベットラウンドが 4 回存在していること，1 ベットラウンドあたりのレイズ回数が無制限であること (クン・ポーカーでは 0 回) に加え，場に全員の共通の手札となるカードが 5 枚存在しており，ベットラウンドが終わるごとにカードの情報が公開されていくことなどから，クン・ポーカーよりも遥かに複雑なゲームである．

4.2 実験方法

第 2.2 節で述べたとおり，2 人零和対称ゲームにおいては，可搾取量が ε である戦略は少なくとも ε-ナッシュ均衡解であることが知られているため，可搾取量が小さいほど良い戦略であると言える．しかしながら，この値は 3 人以上の多人数ゲームにおいては定義されていない．ここで，2 人零和ゲームにおける可搾取量 $\varepsilon_\sigma = b_1(\sigma_2) + b_2(\sigma_1)$ からの類推で，多人数零和ゲームの可搾取量を

$$\varepsilon_\sigma = \sum_{i \in N} b_i(\sigma_{-i})\,, \qquad (17)$$

と定義する．いま，このゲームは零和であるから，各プレ

イヤの利得の和 $\sum_i u_i(\sigma)$ は 0 になる．ゆえに

$$\begin{aligned}\varepsilon_\sigma &= \sum_{i\in N}(b_i(\sigma_{-i}) - u_i(\sigma)) \\ &> b_j(\sigma_{-j}) - u_j(\sigma)\ (\forall j \in N),\end{aligned} \tag{18}$$

が成り立つ ($\forall i \in N,\ b_i(\sigma_{-i}) > u_i(\sigma)$ を用いた)．これを式 (3) に代入することにより，このゲームは少なくとも ε_σ-ナッシュ均衡であることがわかる．従って，本研究では 3 人以上の多人数零和ゲームにおいても，式 (17) を用いて可搾取量を定義することにする．なお，これは Risk らが用いた手法 [14] と本質的に同じものである (式 (17) はゲームが零和であるという条件を用いている点が異なる)．

本稿では FSP の Q 学習，及び教師あり学習には，ソフトマックス回帰 (softmax regression) を用いた．すなわち，確率的勾配降下法によって逐次更新される各アクションについての重みベクトル $\boldsymbol{w}_\Pi(a), \boldsymbol{w}_Q(a)$ があり，ゲームの特徴量ベクトルを $\boldsymbol{x}(s)$ としたとき，教師あり学習は戦略 $\Pi(s,a) = \mathrm{softmax}\left\{\boldsymbol{w}_\Pi(a)^T \boldsymbol{x}(s)\right\}$ を返し，Q 学習は最も適切なアクションとして $Q(s) = \arg\max_a \left\{\boldsymbol{w}_\Pi(a)^T \boldsymbol{x}(s)\right\}$ を返す．それぞれの学習は，NFSP と同様に誤差関数

$$E_\Pi(s,a) = -\log \Pi(s,a), \tag{19}$$

および

$$E_Q(s,a) = \left\{r + \max_{a'} \boldsymbol{w}_\Pi(a')^T \boldsymbol{x}(s') - \boldsymbol{w}_\Pi(a)^T \boldsymbol{x}(s)\right\}^2, \tag{20}$$

を最小化するように学習する．ゲームの特徴量には各プレイヤの手札や掛け額や過去のアクション，それらの直積などを用いた．ハイパーパラメータとして Q 学習のターゲット更新頻度を 400 に，学習率を教師あり学習および Q 学習それぞれで 0.25，1.0 とし，メモリサイズは教師あり学習および Q 学習それぞれで 4×10^4，3×10^3 とした．それ以外は NFSP [2] と同じである．

4.3 実験結果

2 人クン・ポーカーにおける FSP の可搾取量を図 1 に示す．図の横軸 (対数軸) は学習のイテレーション回数であり，縦軸は 1 ゲームあたりの可搾取量である．図 1 からわかるように，可搾取量はイテレーションが進むに従って 0 に近づいているが，0 より少し大きい 0.02 付近に収束している．これは学習部分に単純なソフトマックス回帰を用いたことにより，多層のニューラルネットワークに比べて表現能力が落ちてしまったためであると考えられる．参考のために，図 2 に同一ゲームにおける CFR+の可搾取量を示す．CFR はゲーム木を全探索する方法であるため，2 人クン・ポーカーのような情報集合数の少ないゲームにおいては比較的短い時間で最適戦略を求めることができる．

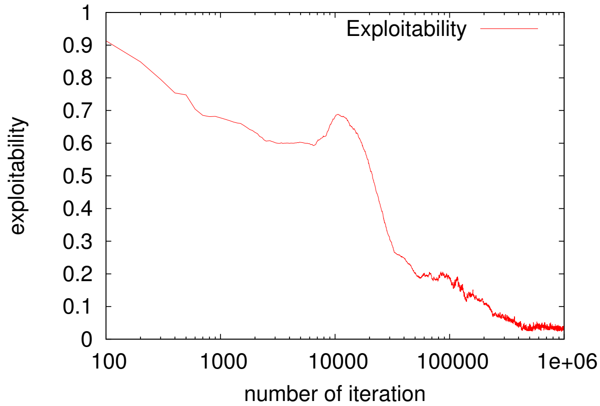

図 1 2 人クン・ポーカーにおける FSP の可搾取量

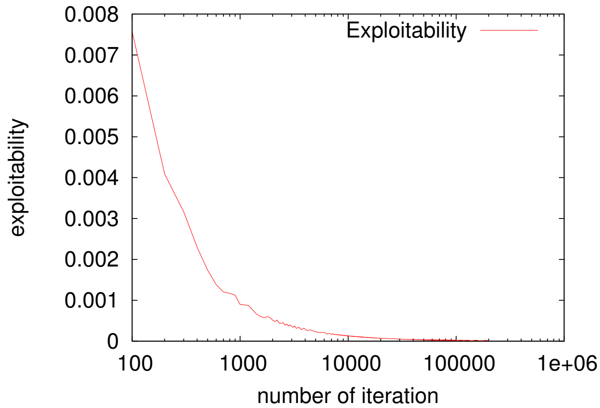

図 2 2 人クン・ポーカーにおける CFR+の可搾取量

3 人クン・ポーカーにおける FSP の可搾取量を図 3 に示す．2 人クン・ポーカーと比べると値は大きくなっているが，イテレーションが進むに従って 0 に近づいていることがわかる．このことから，NFSP には多人数ゲームにおける収束保証性はないが，3 人クン・ポーカーにおいては近似的なナッシュ均衡解に収束すると言える．参考のために，図 4 に同一ゲームにおける CFR+の可搾取量を示す．

3 人クン・ポーカーにおける，FSP と CFR+の対戦で CFR+が得られる期待報酬の推移を図 5 に示す．図の縦軸は，FSP の CFR+に対する平均被搾取量である (可搾取量とは異なることに注意)．図からわかるように，イテレーションが進むに従って FSP のプレイヤが CFR+のプレイヤに搾取される値が減少していることが確認できる．

5. おわりに

5.1 今後の課題

本稿では，テキサス・ホールデムのトイゲームである多人数クン・ポーカーにおいて，ソフトマックス回帰を用い

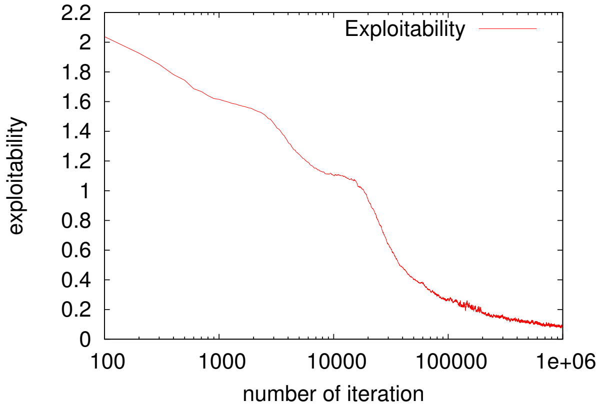

図 3　3 人クン・ポーカーにおける FSP の可搾取量

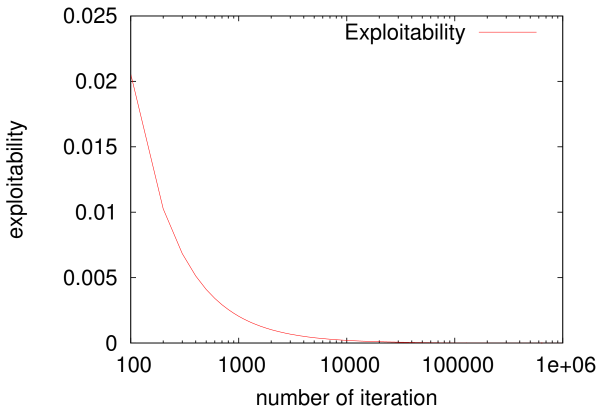

図 4　3 人クン・ポーカーにおける CFR+の可搾取量

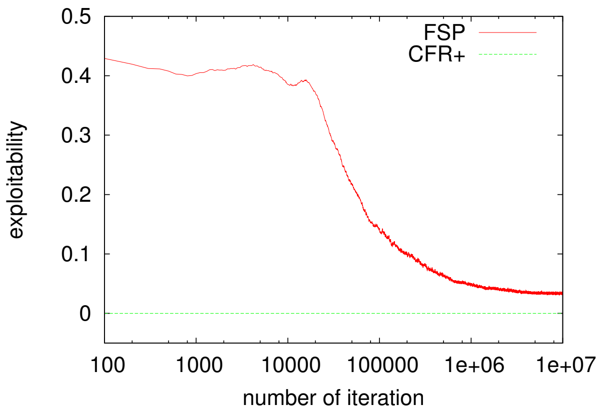

図 5　3 人クン・ポーカーにおける FSP の CFR+に対する平均被搾取量

た FSP の戦略が最適戦略に近づくことを示した．

今後の課題として，学習部分に DQN を使った NFSP を実装することが挙げられる．また，本研究の目的は，CFR が利用できないような状態空間の広い多人数ゲームに対しても NFSP が利用できることを示すことである．従って，3 人以上のテキサス・ホールデムについて，NFSP がナッシュ均衡解に収束するかどうかを実験することが今後の課題である．

参考文献

[1] Durkota, K., Lisỳ, V., Bošanskỳ, B. and Kiekintveld, C.: Optimal network security hardening using attack graph games, *Proceedings of IJCAI*, pp. 7–14 (2015).

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

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

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

[5] 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).

[6] Mnih, V., Kavukcuoglu, K., Silver, D. et al.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, pp. 529–533 (2015).

[7] 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).

[8] 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).

[9] 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).

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

[11] Leslie, D. S. and Collins, E.: Generalised weakened fictitious play, *Games and Economic Behavior*, Vol. 56, No. 2, pp. 285 – 298 (2006).

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

[13] Nisan, N., Roughgarden, T., Tardos, E. and Vazirani, V. V.: *Algorithmic Game Theory* (2007).

[14] Risk, N. A. and Szafron, D.: Using Counterfactual Regret Minimization to Create Competitive Multiplayer Poker Agents, *Proceedings of the 9th AAMAS - Volume 1*, pp. 159–166 (2010).

[15] 古居敬大：相手の抽象化による多人数ポーカーの戦略の決定，修士論文，東京大学大学院工学系研究科電気系工学専攻 (2013).

[16] Shamma, J. S. and Arslan, G.: Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria, *IEEE Transactions on Automatic Control*, Vol. 50, No. 3, pp. 312–327 (2005).

[17] Vitter, J. S.: Random Sampling with a Reservoir, *ACM Transactions on Mathematical Software*, Vol. 11, No. 1, pp. 37–57 (1985).