分散減少法を用いた麻雀における実力推定

大神卓也 *1,a)　天野克敏 *2　奈良亮耶[3]　鶴岡慶雅[1,b)]

概要：麻雀において，プレイヤーの実力を表す指標である平均順位を求める際には，複数の試合の結果を平均するモンテカルロ法が用いられる．この手法では結果の分散が大きく，プレイヤーの実力の信頼区間は幅が大きくなる．そのため，麻雀において正確に実力を推定するためには，膨大な試合数をこなすことが必要になってしまう．例えば，天鳳で十段という素晴らしい成績を達成した Suphx は，十段到達までに約 5 ヶ月かけて 5,760 試合もの試合数をこなしている．そこで本研究では，平均順位に代わって，プレイヤーの実力をよく表す推定値を求めることを目的とする．平均順位の不偏推定量であり，かつ平均順位よりも分散が小さくなる推定値を求めることができれば，同じ幅の信頼区間を求めるために必要な試合数が削減される．同じ不完全情報ゲームであるポーカーでは，対戦データからプレイヤーの実力の推定値を求める研究が行われている．この手法を応用し，深層ニューラルネットワークを導入することと，試合を局単位に分割することによって，ポーカーよりもゲーム木と盤面の情報が大きい麻雀において，推定を行う手法を提案した．そして，インターネット麻雀「天鳳」の牌譜で評価を行ったところ，提案手法を用いて求めた推定値は平均順位と近い期待値をとることが確認できた．しかし，得られた推定値がプレイヤーの平均順位の不偏推定量となっているかについて検定を行ったところ，有意な結果を示すことはできなかった．推定値の分散は平均順位の分散に比べて 14%小さく，プレイヤーの実力を正しく評価するために必要な試合数が 14%減少することが確認された．

Performance Evaluation in Mahjong using a Variance Reduction Technique

Takuya Ogami*[1,a)]　Katsutoshi Amano*[2]　Ryoya Nara[3]　Yoshimasa Tsuruoka[1,b)]

Abstract: In mahjong, the Monte Carlo method, which averages the results of multiple games, is used to find the average rank, a measure of a player's ability. The variance of mahjong results is large, and the confidence intervals for a player's ability are wide. Therefore, a large number of games is required to accurately estimate a player's mahjong ability. For example, Suphx, who achieved an impressive 10-dan in Tenhou, played 5,760 games over a period of about 5 months before reaching 10-dan. Then, the objective of this study is to find an estimate that better represents the players' skill than the average rank. If we can obtain an estimate that is an unbiased estimator of the mean rank and has smaller variance than the mean rank, the number of games needed to obtain a confidence interval of the same width can be reduced. Poker, which is also a game of imperfect information, studies have been conducted to obtain estimates of a player's ability from game data. We proposed a method for estimation in mahjong, where the game tree and information on the board is larger than in poker, by introducing a deep neural network and dividing the game into smaller units. Experiments were conducted on the records on Tenhou, an Internet mahjong game, and it was confirmed that the estimates obtained using the proposed method have an expected value close to the mean rank. However, when we tested whether the obtained estimates were unbiased estimators of the mean rank of the players, the results were not significant. The variance of the estimate is 14% smaller than the variance of the mean rank, confirming that the number of games required to draw the same statistical conclusion is reduced by 14%.

1. 導入

近年では，囲碁 [1] やポーカー [2] など様々なゲームにおいて，人間の実力を超える AI が提案されている．ポーカーや麻雀などの確率的な要素を含むゲームでは，統計的に有意な結論を出すために多くの試合数をこなす必要がある．例えば，インターネット麻雀「天鳳」で十段という素晴らしい成績を達成した Suphx [3] はインターネット麻雀で人間と 5760 試合の対戦を行ったが，それには 4 ヶ月という長い期間を要した．人間と AI の対戦だけでなく，人間どうしの実力の評価も困難である．あるプレイヤーの 1000 試合での平均順位が 2.45 であったとすると，真の平均順位の 95% 信頼区間は 2.45 ± 0.07 となる．従って，1000 試合もの対戦を行ったとしても，真の平均順位の信頼区間は $2.38 \sim 2.52$ となり，大きく勝ち越す実力から，平均的な実力までを含む幅の広い区間となってしまう．そのため，麻雀でプレイヤーが実力を証明するためには膨大な試合数をこなすことが必要である．

ポーカーでは，プレイヤーの実力推定に関する研究が多く行われており [4], [5]，人間のトッププレイヤーと AI との対戦で統計的に有意な結論を導くために活用されている [2]．MIVAT [4] では，ゲーム内でランダムに行われる行動の前後での価値関数の変動を実際の結果から差し引くことによって，プレイヤーの実力の不偏推定量を得ている．AIVAT [5] ではさらに，AI エージェントなどのポリシーが既知のエージェントに対しても同様の手法を適用することによって，プレイヤーの方策が確率論的であることによって生じる分散を削減した．

本研究では，ポーカーよりも複雑なゲームである麻雀において，MIVAT の手法を応用し，プレイヤーが牌山から牌をランダムに引くことによって生じる分散を削減する．ゲームの複雑性に対処するために，価値関数の近似に深層ニューラルネットワークを用い，勾配降下法によって学習を行う．さらに麻雀の巨大なゲーム木を扱うことが難しいため，局と呼ばれるサブゲームに分割を行い，サブゲームごとに利得の予測値を用いて学習を行う．結果として，インターネット麻雀「天鳳」の牌譜で評価を行ったところ，提案手法はモンテカルロ法に比べて分散が 14%小さく，プレイヤーの実力を正しく評価するために必要な試合数が 14% 削減されることが確認された．

[1] 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, Bunkyo, Tokyo 113–8654, Japan
[2] 東京大学大学院情報学環・学際情報学府
Interfaculty Initiative in Information Studies, Graduate School of Interdisciplinary Information Studies, The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[3] 東京大学工学部電子情報工学科
Department of Information and Communication Engineering, The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[a)] ogami@logos.t.u-tokyo.ac.jp
[b)] tsuruoka@logos.t.u-tokyo.ac.jp
[*] These authors contributed equally to this work.

2. 背景

2.1 麻雀の基本的なルール

麻雀は，4 人のプレイヤーが同じ点数を持った状態でゲームが始まり，最終的に 4 人の中で最も高い点数を得ることを目指すゲームである．1 試合は複数の局 (局は試合の小単位) からなる．局において，各プレイヤーのターンで，プレイヤーは山からランダムに 1 枚牌を取る (ツモ) か，特定の条件を満たす場合に相手のプレイヤーが捨てた牌を取る (副露) を行うことができる．そして，和了を宣言するか，手牌の中から 1 枚捨てるかのどちらかを行う．いずれかのプレイヤーが和了形を完成させるか，牌山の牌が無くなった場合に局が終了し，和了形を完成させたプレイヤーは他のプレイヤーから点数を得るため，各プレイヤーは和了形を完成させることを目的として行動する．麻雀牌は 34 種類の牌 4 枚ずつの，合計 136 枚からなる．

2.2 平均順位を用いたプレイヤーの実力推定

麻雀は 4 人で行うゲームであり，ゲーム終了時の各プレイヤーの得点に応じて $1 \sim 4$ 位の順位が定められる．ただし，同じ得点のプレイヤーが複数人存在する場合は，ゲーム開始時のそれぞれの席順に基づいて順位を定める．

あるプレイヤーの真の平均順位の値が μ の時，そのプレイヤーの行った m 試合の結果から μ を推定することで，プレイヤーの実力を評価する．例えば，$1 \sim 4$ 位を同確率で獲得する場合の平均順位 μ は 2.5 であり，あるプレイヤーの平均順位の推定値が 2.5 よりも小さければ，そのプレイヤーは勝ち越しているといえる．

m 試合の結果の平均を $\bar{x}$ とすると，m が十分大きい時，中心極限定理より，$\bar{x}$ は正規分布に近づく．不偏分散 s^2 は式 (1) で表せ，標準誤差 SE は式 (2) となる．

$$s^2 = \frac{1}{m-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1}$$

$$\mathrm{SE} = \frac{s}{\sqrt{m}} \tag{2}$$

この時，母平均の 95%信頼区間は式 (3) で求められる．

$$\bar{x} - 1.96\mathrm{SE} \leq \mu \leq \bar{x} + 1.96\mathrm{SE} \tag{3}$$

SE を一定に保つ時，試合結果の分散 s^2 を削減できれば，評価に必要な試合数 m も削減することができる．

2.3 展開型ゲーム

展開型ゲーム [6] は複数人での不完全情報を伴う環境での意思決定を扱うための枠組みである．展開型ゲーム

は $\left(\mathcal{N} \cup \{c\}, \mathcal{H}, \mathcal{Z}, \mathcal{A}, \{R^i\}_{i \in \mathcal{N}}, \mathcal{P}, \pi^c, \mathcal{S}\right)$ によって定義される．$\mathcal{N}$ は N 人のプレイヤーの集合 $\{1, \ldots N\}$ を表す．c はチャンスプレイヤーと呼ばれ，環境のランダム性を表現するために導入されるプレイヤーである．例えば，ポーカーにおいて最初に配られる手札はチャンスプレイヤー c の行動の結果として表現することができる．チャンスプレイヤーは方策 π^c に従って行動する．履歴 $h \in \mathcal{H}$ はゲーム開始時からの全てのプレイヤーの行動の系列を表し，$\mathcal{H}$ は取りうるすべての履歴の集合を表す．履歴 $h_1, h_2 \in \mathcal{H}$ について，h_1 が h_2 の行動系列の prefix であるか，$h_1 = h_2$ であるとき，$h_1 \sqsubseteq h_2$ と表す. 履歴 h においてプレイヤーが行動 a を行って得られる履歴を $h \cdot a$ と表す．終端履歴集合 $\mathcal{Z} \subseteq \mathcal{H}$ は履歴のうち，ゲームの終了時点を表す履歴の集合を表す．$\mathcal{A}(h)$ は非終端履歴 $h \in \mathcal{H} \setminus \mathcal{Z}$ において取りうる行動の集合を表す．各エージェントは終端履歴において，エージェント $i \in \mathcal{N}$ に対する利得関数 $R^i : \mathcal{Z} \to \mathbb{R}$ に従って利得を得る．$\mathcal{P} : \mathcal{H} \to \mathcal{N} \cup \{c\}$ は履歴において行動するプレイヤーを表す関数である．履歴集合は履歴において行動を行うプレイヤーからの見え方が同一なものをまとめることによって，$s_1, s_2, \ldots$ のように分けられる．$\{s_1, s_2, \ldots\}$ を情報集合 $\mathcal{S}$ とする．

2.4 MIVAT

MIVAT [4] は，展開型ゲームにおいて分散を削減する手法である．MIVAT ではプレイヤーの対戦のデータから，プレイヤーの利得の不偏推定量であり，利得よりも分散が小さいような推定量を求める．式 (4) のように，プレイヤーが運よく得た利得 l(「運」と呼ぶ) をプレイヤーが得た利得 u から差し引いた値 $\hat{u}$ を利得の推定値とする．

$$\hat{u} = u - l \tag{4}$$

式 (5) のように,「運」の期待値が 0 であれば，推定値 $\hat{u}$ は利得 u の不偏推定量になる．

$$E[\hat{u}] = E[u - l] = E[u] \tag{5}$$

式 (6) のように, プレイヤー j の運 L_j を価値関数 $V_j : \mathcal{H} \to \mathbb{R}$ を用いて計算する．終端履歴 z をとったゲームでのプレイヤー j の運は，履歴に含まれるチャンスプレイヤーの行動前後の価値関数の変動の総和となる．

$$L_j(z) = \sum_{\substack{h \ s.t. \\ h \cdot a \sqsubseteq z, P(h) = c}} \left(V_j(h \cdot a) - V_j(h) \right) \tag{6}$$

$$= \sum_{\substack{h \ s.t. \\ h \cdot a \sqsubseteq z, P(h) = c}} \left(V_j(h \cdot a) - \sum_{a'} \pi^c(a'|h) V_j(h \cdot a') \right) \tag{7}$$

ここで，式 (7) のように $V_j(h)$ を展開することで，任意の価値関数 $V_j : \mathcal{H} \to \mathbb{R}$ に対して，$L_j(z)$ の期待値が 0 となる．価値関数の学習として入力特徴量の線形関数を用いる．式 (8) の最小化を目的関数とすることで，線形関数の最適なパラメータを解析的に求めることができる．

$$\underset{V_j}{\text{Minimize}} \ \text{Var}(\hat{u}) \tag{8}$$

ポーカーや麻雀などの N 人ゼロサムゲーム ($N > 2$) については，$\sum_{j=1}^{N} V_j(h) = 0$ となるように制約を設ける．

3. 関連研究

3.1 麻雀における実力推定

久米らの研究 [7] では式 (9) で表されるエラーレートを用いて麻雀プレイヤーの実力を推定した．エラーレートの計算はプレイヤーの状態行動対 (s, a) の集合 $\mathcal{D}$ に対し，AI プレイヤーの状態価値関数 $Q(s, a)$ を用いて式 (9) のように計算される．$Q(s, a)$ を基準とし，プレイヤーのとった選択の状態価値関数の値が小さいほど，エラーレートが大きくなる．

$$e(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(s,a) \in \mathcal{D}} \left(\max_{a' \in A(s)} Q(s, a') - Q(s, a) \right) \tag{9}$$

状態価値関数には天鳳七段を達成した AI プレイヤーである「Ako_Atarashi」の価値推定の結果を用いている．

久米らはインターネット麻雀「天鳳」における 500 試合のレートを推定する際に，500 試合の平均順位を用いて推定した結果と，16 試合程度の試合で推定した結果の精度が同程度であると主張した．推定値が求めたい変数の不偏推定量になっているわけではないという点，特定の AI の価値関数に依存するという点が本研究とは異なっている．

3.2 不完全情報ゲームでの分散削減

Bowling ら [8] は重点サンプリングを用いて実際に行われたゲームの途中でゲームを終了させるような行動をとる場合を考慮すること, プレイヤーの private な情報以外が同一の局面（すなわち，相手の方策は実際のゲームと変わらない）を考慮することによって分散を削減した．Action-Informed Value Assessment Tool(AIVAT) [5] では MIVAT [4] を拡張し，ポーカーにおける結果の分散を削減するために，(1) 配られるカードなどのゲーム内で確率的に決定される要素に起因する分散，(2) プレイヤーの方策が確率論的であることによって生じる分散の両方を削減した．DeepStack [2] で人間との対戦結果の評価の際に使用され，分散を 85%削減し評価に必要な試合数を 44 分の 1 まで削減した．

麻雀では決定論的な方策を考慮することが多いため，本研究では (2) の確率論的な方策に起因する分散については考慮していない．また，ポーカーよりも状態数が多く，ゲームの履歴が長い麻雀という問題に取り組んでいる．

3.3 Off-Policy Evaluation における分散削減

ある方策によって得られた経験から別の方策における価値関数を決定する問題は Off-Policy Evaluation と呼ばれている．Off-Policy Evaluation では重点サンプリングを用いる手法 [9] が代表的であるが，重点サンプリングを用いると分散が大きくなってしまうことが知られている．Jiang ら [10] は，強化学習における Off-Policy Evaluation において，DoublyRobust 法 (DR 法) を組み合わせて価値関数を構成することで，不偏推定かつ分散が小さい価値関数を構築する手法を提案した．

経験における方策と評価したい方策が一致している点，削減したい分散が重点サンプリングに由来するものではなく，ゲームの運の要素に由来するものである点が本研究とは異なる．

4. 提案手法

4.1 麻雀におけるチャンスプレイヤーの行動

麻雀におけるチャンスプレイヤーの行動は (1) プレイヤーが各手番で牌山からランダムに 1 枚牌を引く (2) 局のはじめに配牌がランダムに配られる (3) 表ドラ，裏ドラがランダムに決定される の 3 種類である．そこで本稿では (1) のプレイヤーがランダムに牌山から 1 枚引く行動のみを取り扱った．配牌やドラに関するチャンスプレイヤーの行動を考慮することによって，更なる分散の削減が期待される．

4.2 MIVAT に対する深層ニューラルネットワークの導入

MIVAT [4] では価値関数の近似に線形関数を用いて, 標本分散を最小化することによって解析的に価値関数を求めた．一方で, Suphx [3] をはじめとした麻雀対戦用 AI では上級者の模倣を行うために Convolutional Neural Network(CNN) を用いている．本研究でも麻雀の盤面の複雑な情報を効果的に処理するために，価値関数の近似に線形関数の代わりに CNN を用いる．価値関数の最適化を行うために，確率的勾配降下法のアルゴリズムを用いる．Algorithm 1 に本研究で用いたアルゴリズムを示している．提案手法ではミニバッチごとに $\hat{u}$ の不偏分散を計算し誤差関数として用いている．誤差関数の計算を安定させるために，バッチサイズを可能な限り大きく設定する．

4.3 Global Reward Prediction(GRP) によるゲームの分割

2.1 節で述べた通り，麻雀の一試合は局と呼ばれるサブゲームを繰り返す. 麻雀のゲーム木は大変大きく，試合の最後にしか利得が与えられないため，学習が困難である．局終了時にプレイヤー間の得点の移動が生じるが，それが試合最後の利得にどう影響するか自明ではない．そこで本研究では，麻雀の一試合を局の単位で分割し，それぞれの局の終了時の利得を推定する．この工夫は Suphx [3] において，強化学習の報酬の定義で用いられていたものである．

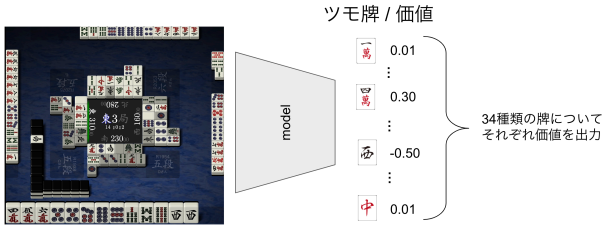

図 1 価値関数の概略図。プレイヤーがツモを行う直前の場面を入力とし，34 種類の牌それぞれをツモった場合の価値関数を出力する．

試合終了時に実際に獲得できた順位を $x \in \{1,2,3,4\}$ とし，k 局目終了時の状態から，その試合終了時に獲得できると予測した順位を $\hat{x}_k \in [1,4]$ とする．ただし，$\hat{x}_0 = 2.5, \hat{x}_K = x$ である．このとき，局 $k = 1, \dots K$ での利得は $u_k = \hat{x}_k - \hat{x}_{k-1}$ で定義できる．

u_k に対し MIVAT を適用して分散を削減した推定値を $\hat{u}_k$ とすると，獲得順位の推定値 y は，それぞれの局の終了時の利得の推定値 $\hat{u}_k$ の和で表せ，

$$\begin{aligned} y &= \sum_{k=1}^{K} \hat{u}_k \\ &= \sum_{k=1}^{K} u_k - L_k \\ &= \sum_{k=1}^{K} \hat{x}_k - \hat{x}_{k-1} - L_k \\ &= x - 2.5 - \sum_{k=1}^{K} L_k \end{aligned} \tag{10}$$

となる．したがって，(獲得順位の推定値 y) = (実際に獲得した順位 x) - (各局での「運」の和) が成り立ち，局に分割して学習を行うことができる．

5. 実験

5.1 データセット

インターネット麻雀「天鳳」[1]の最上位のフィールドである，鳳凰卓の牌譜を用いる．訓練，開発セットとして 2020 年の牌譜のうち，およそ 400,000 局のデータを用いる．テストには，以下の 2 種類のデータセットを用意する．

プレイヤー混合データセット
: 2021 年の牌譜のうち，33,664 試合のデータを用いたデータセット．

プレイヤー別データセット
: 2021 年の牌譜のうち，56 名のプレイヤーごとに，それぞれ 500 試合以上のデータを用いたデータセット．

[1] https://tenhou.net/

Algorithm 1 DNN と MIVAT の学習

Input: $V_{\boldsymbol{\theta}}^{p}$; value function of player $p(p = 1, 2, 3, 4)$, N :number of iteration for each epoch, B: batch size z: terminal history, $\boldsymbol{u} = \{u_1, u_2, u_3, u_4\}$: utility for each player

for $t = 1, \cdots, N$ **do**
  sample B tuples of (z, $\boldsymbol{u}$)
  initalize $\boldsymbol{Luck} = \{0, 0, 0, 0\}$
  for all history_before_chance_action$h \sqsubseteq z$ **do**
    $V^{p}(h, i)_{i=\{1,\cdots 34\}} = V_{\theta}^{p}(h)$
    weight $V^{p}(h, i)$ according to equation (7) and add it to $\boldsymbol{Luck}$
  end for
  $\hat{\boldsymbol{u}} = \boldsymbol{u}$ - $\boldsymbol{Luck}$
  L = sample_variance($\hat{\boldsymbol{u}}$)
  update $\boldsymbol{\theta}$ using Adam[11]
end for

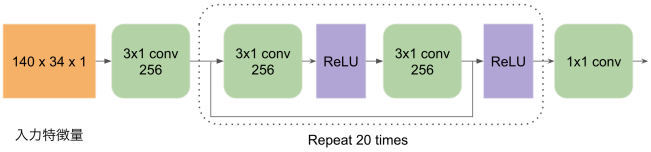

図 2 価値関数の近似に用いるニューラルネットワーク．Suphx [3] の打牌選択に用いられたモデルと同様の構造を用いているが，より小規模なモデルを用いている点が異なる．

データセットは局ごとに分割され，4.3 で述べたように Global Reward Prediction を用いて利得の予測値をラベルとして付与する．

5.2 モデルの設定

価値関数の近似には，Suphx [3] で利用された残差接続 [12] を用いた深層ニューラルネットワークを参考に実装する (図 2). ハイパーパラメータの設定は表 1 に記載した通りである．モデルの特徴量を表 2 に示す．価値関数の入力は展開型ゲームにおける履歴であるため，不完全情報も入力に含まれる．Suphx [3] や Variational Latent Oracle Guiding (VLOG) [13] では麻雀において不完全情報をエージェントに利用させるための研究を行った．Suphx では相手の手牌と牌山を，VLOG では相手の手牌のみを，Oracle 情報として入力している．本研究では特徴量を減らすため，VLOG にならって相手の手牌のみを特徴量として採用する．

パラメータ	値
最適化手法	Adam [11]
学習率	0.0001
バッチサイズ	150

表 1 モデルの学習に関する設定

特徴量	チャネル数
親	4
自風	4
場風	4
本場	6
供託	6
保持点数	80
ドラ	8
自分の手牌	7
相手の手牌	21
合計	140

表 2 盤面の情報を表すための特徴量の詳細.

5.3 評価指標

5.3.1 プレイヤー混合データセットを用いた評価

テストデータ M 試合において，実際に獲得した順位を $\{x_i\}_{i=1}^{M}$，提案手法によって推定した利得を $\{y_i\}_{i=1}^{M}$ とする．x, y の標本平均を $\bar{x}, \bar{y}$，不偏分散を s_x^2, s_y^2，母平均を μ_x, μ_y とする．提案手法によって，(1) 分散が削減できたか，(2) 不偏推定が行えているか，を評価する．

(1) 分散の評価では，s_x^2 と s_y^2 から，x, y の母分散の等しさについて，F 分布を用いて検定を行う．x, y の母分散を σ_x^2, σ_y^2 とする．帰無仮説 $H_0 : \sigma_x^2 = \sigma_y^2$，対立仮説 $H_1 : \sigma_x^2 > \sigma_y^2$ として，有意水準 5% で片側検定を行う．式 (11) で表される x, y の不偏分散の比 f は自由度 $(M-1, M-1)$ の F 分布に従う．下位 5% 点 $F_{0.05}(M-1, M-1)$ を求め，$f < F_{0.05}(M-1, M-1)$ の時，帰無仮説が棄却され，対立仮説が採択される．すなわち，推定値 y は実際に獲得した順位 x よりも有意に分散が小さいことが言える．

$$f = \frac{s_y^2}{s_x^2} \tag{11}$$

(2) 不偏推定の評価では，x と y の母平均の差が，ある閾値 $\delta > 0$ よりも小さいことの検定を行う．$\bar{x} > \bar{y}$ と $\bar{x} < \bar{y}$ の両方の場合について検定を行う必要がある．

まず $\bar{x} > \bar{y}$ の場合，帰無仮説 $H_0 : \mu_x - \mu_y = \delta$，対立仮説 $H_1 : \mu_x - \mu_y < \delta$ として，有意水準 5%で片側検定を行う．試合数 M が十分大きい時，μ_x, μ_y は正規分布に従うと近似できるため，式 (12) で示された統計検定量 z_1 は標準正規分布に従う．したがって $z_1 < -1.64$ のとき，対立仮説が採択され，$\mu_x - \mu_y$ は十分小さいと言える．

一方，$\bar{x} < \bar{y}$ の場合には，帰無仮説 $H_0 : \mu_y - \mu_x = \delta$，対立仮説 $H_1 : \mu_y - \mu_x < \delta$ として，検定を行う．式 (13) で示された z_2 に対して，$z_2 < -1.64$ のとき，対立仮説が採択され，$\mu_y - \mu_x$ は十分小さいと言える．

$$z_1 = \frac{(\bar{x} - \bar{y}) - \delta}{\sqrt{\frac{s_x^2 + s_y^2}{M}}} \quad (\bar{x} > \bar{y}) \tag{12}$$

$$z_2 = \frac{(\bar{y} - \bar{x}) - \delta}{\sqrt{\frac{s_x^2 + s_y^2}{M}}} \quad (\bar{x} < \bar{y}) \tag{13}$$

5.3.2 プレイヤー別データセットを用いた評価

データセットには様々な実力のプレイヤーのデータが含まれており，プレイヤーごとに獲得順位の母平均が異なる．上述のプレイヤー混合データセットではそれを考慮できていない．そこで，プレイヤー別にデータセットを作成し，評価を行う．テストデータに含まれるプレイヤー数を n とし，テストデータ $M = \sum_{j=1}^{n} m_j$ 試合において，プレイヤー j が実際に獲得した順位の集合をを $\mathcal{X}_j = \{x_{i,j}\}_{i=1}^{m_j}$，提案手法によって推定した利得を $\mathcal{Y}_j = \{y_{i,j}\}_{i=1}^{m_j}$ とする．$\mathcal{X}_j, \mathcal{Y}_j$ の標本平均を $\bar{x}_j, \bar{y}_j$，不偏分散を $s_{x,j}^2, s_{y,j}^2$，母平均を $\mu_{x,j}, \mu_{y,j}$ とする．

推定値の不偏性は，各 j に対し，式 (12)，(13) を求めることで検証できる．不偏分散の減少の程度を表す指標として，式 (14) を導入する．各プレイヤーの分散の減少割合を，プレイヤーの評価試合数によって重み付けして加算した指標である．

$$\sum_{j=1}^{n} \frac{m_j}{M} \frac{s_{y,j}^2}{s_{x,j}^2} \tag{14}$$

6. 結果

開発セットでの利得の分散は 0.119 であった．モデルの訓練を行ったところ，開発セットにおける推定値の分散は 0.0961 まで減少し，利得の分散に比べて小さくなった．

6.1 プレイヤーを区別しない評価

プレイヤー混合データセット 33,664 試合における獲得順位の平均と標準誤差を表 3 に示す．提案手法における不偏分散は 1.0673 であり，実際の獲得順位（モンテカルロ法）における不偏分散 1.2386 と比較して約 14% 減少した．また，式 (11) に基づいて計算すると，$f = 0.8617$ であった．$M = 33,664$ のとき，$F_{0.05}(M-1, M-1) = 0.9822$ であり，$f < F_{0.05}(M-1, M-1)$ が成り立つ．すなわち，帰無仮説 $H_0 : \sigma_x^2 = \sigma_y^2$ が棄却され，有意水準 5% の下で有意に分散が小さくなった．

続いて，不偏推定ができたかについて評価を行う．標本平均の差 $\bar{x} - \bar{y} > 0$ より，統計検定量は式 (12) に基づいて求める．$\delta = (0.01, 0.02, 0.03)$ でそれぞれ求めたところ，$z_1 = (-1.159, -2.367, -3.576)$ となった．従って，$\delta = 0.02$ の元では，$z_1 < -1.64$ より，帰無仮説 H_0 が棄却され，対立仮説 H_1 が採択される．すなわち，有意水準 5%，$\delta = 0.02$ の範囲で不偏推定が成り立っているといえる．$\delta = 0.02$ は，標本平均の差 $\bar{y} - \bar{x} = 0.0004$ と比較すると，大きい値となってしまっている．そのため，評価試合数を増やし，さらに厳密な不偏性が成り立つか検証する必要がある．

6.2 プレイヤーを区別した評価

プレイヤー別データセットにおいて式 (14) を求めたところ，0.8533 となり，提案手法によって分散がおよそ 14% 削減されたことが確認された．図 3 では，評価対象のプレイヤー j ごとに値をプロットしている．縦軸は，m_j 試合における獲得順位の不偏分散 $s_{x,j}^2$ と，推定利得の不偏分散 $s_{y,j}^2$ の比 f を示す．全てのプレイヤーに対し，$f < 1$ となっており，提案手法によって不偏分散が削減されていることがわかる．また，横軸は $F(m_j - 1, m_j - 1)$ の下位 5% 点を示し，赤線は $f = F_{0.05}(m_j - 1, m_j - 1)$ を表す．等分散性の検定を行うと，赤線より右下側にプロットされたプレイヤーに関して，帰無仮説 $H_0 : \sigma_x^2 = \sigma_y^2$ が棄却される．従って，56 人中 8 人を除く 48 人のプレイヤーについて，有意水準 5% の下で有意に分散が削減できたことが示された．

また，図 4 では評価対象の 56 名のプレイヤーごとに，式 (12)，(13) に基づく統計検定量 z_1, z_2 をプロットしている．ただし，閾値は $\delta = 0.08$ とした．図中の赤線より下の領域にプロットされているプレイヤーに関しては，帰無仮説 H_1 が棄却され，有意水準 5% の範囲で不偏推定になっているといえる．しかし，今回の実験では，帰無仮説が棄却される領域にプロットされたのは 56 人中 8 人にとどまり，個別のプレイヤーについて，今回求めた推定値が平均順位の不偏推定となっているとはいえない．原因としては，評価に用いる試合数が不十分であることが考えられる．プレイヤーごとのデータセットでは各プレイヤーごとの試合数が 2,500 件程度に限られており，統計的に有意な結論を出すためには十分ではない．

	平均	不偏分散
モンテカルロ法	2.4955	1.2386
提案手法	2.4959	1.0673

表 3 プレイヤー混合データセット 33,664 試合における獲得順位と推定利得の平均・分散

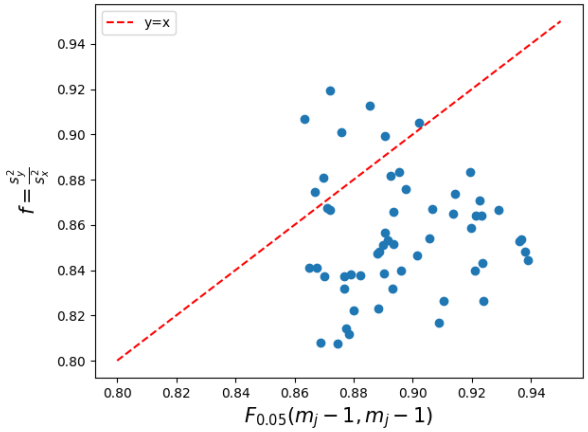

図 3 評価対象のプレイヤーごとに，式 (11) に基づく f と，自由度 $(m_j - 1, m_j - 1)$ の F 分布の下位 5% 点を図示．

7. おわりに

本稿では麻雀において，プレイヤーの実力を推定する手法について検討した．ポーカーで主に用いられている，不

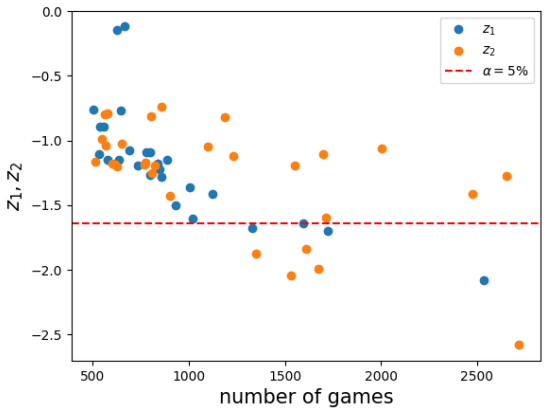

図 4 評価対象のプレイヤーごとに，$\bar{x} > \bar{y}$ のプレイヤーは統計検定量 z_1 を，$\bar{x} < \bar{y}$ のプレイヤーは z_2 を図示．($\delta = 0.08$)

完全情報ゲームにおいてプレイヤーの利得の期待値に関する推定を行う MIVAT [4] を，より盤面の情報が複雑でゲーム木が大きい麻雀に適用する手法について提案した．提案手法を用いることによって分散が減少し，期待値が標本平均の期待値と近いような推定値を求められることが確認できた．一方で，得られた推定値がプレイヤーの平均順位の不偏推定量となっているかについて検定を行ったところ，有意な結果を示すことはできなかった．さらに，分散の減少度もおよそ 14%にとどまる結果となった．

今後の展望としては，データセットとして使用するデータの数を増加させること，分散削減の対象となるイベントを増加させることが挙げられる．本稿では，訓練用のデータセットとしておよそ 400,000 局分のデータセットを用いたが，一局あたりにチャンスプレイヤーが行動を行う場面はとても多いため，それぞれのイベントが最終的な報酬にどの程度寄与するかを推測するためには大規模なデータが必要になると考えられる．また，推定値の不偏性を評価する際にもより多くのデータを用いることが望ましい．

4.1 節で述べた通り，本稿では分散削減の対象として，プレイヤーのツモの前後の価値関数の変動にのみ着目している．プレイヤーの配牌前後の価値関数の変動を考慮することによって分散の削減度を向上させることができると考えている．

参考文献

[1] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, Vol. 529, No. 7587, pp. 484–489, 2016.

[2] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, Vol. 356, No. 6337, pp. 508–513, 2017.

[3] Junjie Li, Sotetsu Koyamada, Qiwei Ye, Guoqing Liu, Chao Wang, Ruihan Yang, Li Zhao, Tao Qin, Tie-Yan Liu, and Hsiao-Wuen Hon. Suphx: Mastering mahjong with deep reinforcement learning. *arXiv preprint arXiv:2003.13590*, 2020.

[4] Martha White and Michael H Bowling. Learning a value analysis tool for agent evaluation. In *IJCAI*, pp. 1976–1981, 2009.

[5] Neil Burch, Martin Schmid, Matej Moravcik, Dustin Morrill, and Michael Bowling. Aivat: A new variance reduction technique for agent evaluation in imperfect information games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, 2018.

[6] Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

[7] 久米洋輝, 栗田萌, 保木邦仁ほか. エラーレートを用いた麻雀プレイヤの実力推定. 研究報告ゲーム情報学 (GI), Vol. 2019, No. 14, pp. 1–7, 2019.

[8] Michael Bowling, Michael Johanson, Neil Burch, and Duane Szafron. Strategy evaluation in extensive games with importance sampling. In *Proceedings of the 25th international conference on Machine learning*, pp. 72–79, 2008.

[9] T. Mandel, Y.-E Liu, S. Levine, E. Brunskill, and Z. Popović. Offline policy evaluation across representations with applications to educational games. *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, Vol. 2, pp. 1077–1084, 01 2014.

[10] Lihong Li Nan Jiang. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of The 33rd international conference on Machine Learning*, pp. 652–661.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[12] S Jian, H Kaiming, R Shaoqing, and Z Xiangyu. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 770–778, 2016.

[13] Dongqi Han, Tadashi Kozuno, Xufang Luo, Zhao-Yun Chen, Kenji Doya, Yuqing Yang, and Dongsheng Li. Variational oracle guiding for reinforcement learning. In *International Conference on Learning Representations*, 2021.