修 士 論 文

強化学習を用いたコンピュータ麻雀プレイヤ

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

山田 渉央

2018 年 3 月

修 士 論 文

強化学習を用いたコンピュータ麻雀プレイヤ

指導教官　池田 心

審査委員主査　池田 心
審査委員　飯田 弘之
審査委員　白井 清昭

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

1510056 山田 渉央

提出年月: 2018 年 2 月

概 要

ゲームの相手をコンピュータプレイヤが行って人間と対戦する場合，人間プレイヤが楽しいと思えることが一般的な目標と言える．そのための条件はいくつもあるが，多くの人間プレイヤは，まず相手が互角程度の実力を持つことを要求すると考える．十〜数十年前であれば互角レベルのプレイヤを作ることすら難しいゲームが多かったため，様々なジャンルのゲームで強いコンピュータプレイヤを作るための研究が盛んに行われてきた．

チェスでは，1997 年に IBM 社の Deep Blue が世界チャンピオンに勝利し，将棋においても将棋電脳戦などでプロ棋士と互角以上の実力を示している．囲碁においては 2017 年に DeepMind 社が開発した AlphaGo が人間の世界トッププレイヤに勝利している．チェスや将棋，囲碁はゲームの盤面の状態がプレイしている 2 人のプレイヤに公開されている完全情報ゲームであり，プレイヤのどちらかが勝利した場合，もう一方のプレイヤは必ず敗北となるゼロサムゲームである．このジャンルのゲームにおいては，すでに人間のトップレベルのプレイヤと同等かそれ以上の実力を持つコンピュータプレイヤが作られたと言える．そこで，強いコンピュータプレイヤを作る研究はこれらよりもなんらかの意味で複雑な要素を持つゲームへと対象が移っている．その 1 つとして麻雀が挙げられる．

麻雀は 4 人で行われる不完全情報ゲームである．プレイ人数が多い点，ランダム性がある点，不完全な情報を推測する必要がある点，そして 1 回の局の勝敗が直接ゲームの勝敗になるわけではない繰り返しゲームである点などが，チェスなどに比べて複雑な要素だと思われる．麻雀における意思決定の難しさの 1 つとして，手牌や捨て牌などの状態が同じでも，現在の順位や点数，残りゲーム数などに応じて取るべき戦略が異なる点が挙げられる．例えば，現在の順位が高い場合は，期待される上がり点数が低くなったとしてもゲームを早く終わらせる戦略が有効となる場合が多い．一方で，現在の順位が低く，上位の選手との点数差が大きい場合は，上がるのが難しい点数の高い手を作る戦略が有効となる場合が多い．どちらにも偏らない中間の戦略も考えられる．このように，状況に応じた多様な戦略が必要となること，また麻雀は利用されるルールが様々でそれぞれ適切な戦略が異なることなど，これらは人の手によるコンピュータの戦略作成を困難にしている．

そこで本研究では，これらの異なる戦略をできるだけ容易に得るため，強化学習による麻雀プレイヤの作成を提案する．強化学習の報酬や設定を変えることで異なる目的や傾向を持った戦略を自動的に獲得することを目指す．実験では，典型的に異なる 2 つの戦略を得るため，“ 上がった場合に固定の報酬を得る学習 ”と“ 上がった時の和了点数を報酬とする学習 ”を行った．一定報酬の場合は，アガリまでの平均手数が少なくなるようなプレイヤが，点数報酬の場合は上がった時の平均点数が高くなるようなプレイヤが得られると期待できる．また割引率パラメータの調整などによりこれらの中間的なものも得られると期待できる．

はじめに,「実験が高速であること」「理論値の計算と比較ができること」を理由に，麻雀を単純化した小規模なゲームに強化学習を適用した．この単純化ゲームは麻雀として最低限の要素を残しつつ，ルールを簡単にし，状態空間の大きさを小さくしたものである．本研究では 5 種類の単純化ゲームを設計し，複数の強化学習手法が問題の規模に応じてどのような得手不得手を持つのかを確認した．

単純化ゲームのうち 3 枚麻雀，5 枚麻雀については手法 1 つ目，全状態行動価値（Q 値）をテーブルに保存し，それを更新する“ テーブル型 ”の実験を行った．テーブル型は それぞれの価値を個別に更新することができるため，充分に学習が進めば理論値通りの価値を得られることが利点である．一方で，状態の数が多くなるとメモリに全状態を載せられない，載せられたとしてもその状態を経験するのに時間がかかるという欠点がある．2 色 5 枚麻雀では 10 万通り程度の状態であるためテーブル型での実験が可能であるが，2 色 8 枚麻雀になると困難であった．

手法２つ目として，手牌を特徴量行列で表し，その次元数と同じ要素数を持つ重みベクトルとの積で状態行動価値を表現する“特徴量型”という手法を用いた．特徴量の次元数は 5 枚麻雀で 76，8 枚麻雀では 112 とした．特徴量型の学習はメモリに乗せることが容易になる反面，手牌ごとに個別に価値の更新ができないため，理論値通りの結果が得られない可能性があり，テーブル型よりも限界性能は悪くなると予想される．5 枚麻雀でテーブル型，特徴量型の上がるまでの平均手数を比較したところ，それぞれ 11.9 手，12.4 手となり，予想の通りテーブル型の方が性能は良かった．

続いて，実際の麻雀のルールを用いた 1 人麻雀に対して強化学習を適用した．特徴量の次元数は 192 とした．テーブル型は用いることができないので，特徴量型の強化学習を用い，“一定報酬による平均手数の少ないプレイヤ”と，“獲得点数報酬による上がった時の平均点数が高くなるプレイヤ”を実現した．具体的には，それぞれ 30.3 手 3300 点と，34.8 手 7700 点であり，その個性は顕著である．また，獲得点数報酬を与えつつ，早い上がりを誘導するために割引率パラメータを 0.9 から 0.7 に下げることにより，32.6 手 4200 点と，中間の戦略を取るプレイヤを実現した．

最後に手法３つ目として，多層ニューラルネットワークを用いて状態行動価値を表現する“ニューラルネット型”の手法を実装した．これにより，前述の特徴量型に比べて，状態行動価値に対する表現力が上がり，学習性能の向上が期待される．一方で，特徴量型に比べて計算式が複雑になるため同じ試行回数ではより計算に時間が掛かることがデメリットとして考えられる．3 色 8 枚麻雀に適用したところ，試行数の少ない段階では，特徴量型に比べてアガリ確率が向上するまでに必要な試行回数が少なくなることを確認した．例えば，90 %のアガリ確率を得るまでに必要な試行回数は約 60 万から約 3 万に減らすことができた（ただし，実行時間はさほど変わらない）．

目 次

第 1 章　はじめに **1**
1.1　研究背景 1

第 2 章　対象ゲーム **3**
2.1　麻雀のルール 3
2.2　麻雀における戦略 5
2.3　1 人麻雀 9

第 3 章　関連研究 **10**
3.1　牌譜を用いた教師あり学習による 1 人麻雀プレイヤ 10
3.2　複数の予測モデルを組み合わせた麻雀プレイヤ 11
3.3　モンテカルロ法を応用した一人麻雀プレイヤ 12
3.4　麻雀の複数の戦略に対する着手モデル 13
3.5　麻雀への多層ニューラルネットワークの適用 14
3.6　不完全情報ゲームにおける多層ニューラルネットワークによる強化学習の価値関数の近似 . 15

第 4 章　強化学習 **16**
4.1　テーブル型 16
4.2　特徴量型 17
4.3　ニューラルネット型 18
4.4　学習パラメータと得られる戦略 19
4.4.1　割引率 19
4.4.2　報酬 20

第 5 章　単純化ゲーム **21**
5.1　3 枚麻雀 22
5.1.1　3 枚麻雀のルール 22
5.1.2　3 枚麻雀のテーブル型実験 22
5.1.3　3 枚麻雀の特徴量型実験 25
5.2　1 色 5 枚麻雀 27
5.2.1　1 色 5 枚麻雀のルール 27
5.2.2　1 色 5 枚麻雀実験 27
5.3　2 色 5 枚麻雀 31
5.3.1　2 色 5 枚麻雀のルール 31
5.3.2　2 色 5 枚麻雀実験 31
5.4　2 色 8 枚麻雀 34

5.5 3 色 8 枚麻雀 37
5.5.1 特徴量型実験 37
5.6 単純化ゲーム問題まとめ 39

第 6 章 1 人麻雀 **40**
6.1 特徴量型実験 40
6.1.1 実験設定 40
6.1.2 実験結果 1 43
6.1.3 具体的な戦略例 45
6.1.4 既存研究との比較 46
6.2 手数グルーピング手法実験 47
6.3 ニューラルネット型 49
6.3.1 実験の目的 49
6.3.2 実験設定 49
6.3.3 実験結果 50

第 7 章 まとめ **52**

第1章　はじめに

1.1　研究背景

コンピュータプレイヤが人間と対戦する際，人間プレイヤが楽しいと思えるということが重要であると考える．そのための条件はいくつも考えられるが，多くの人間プレイヤは相手が互角程度の実力であることを要求すると思われる．そのために，様々なジャンルのゲームで強いコンピュータプレイヤの研究が盛んにおこなわれてきた．チェスや将棋，囲碁においては人間のトッププレイヤに勝利するなど [1] [2] [3] 十分に強くなったと言え，面白さについての研究が行われるようになった [4]．そこで，強いコンピュータプレイヤについての研究は，これらのゲームに比べて何らかの難しさを持つゲームへと対象が移ってきている．その 1 つとして麻雀が挙げられる．国内における愛好者が多い麻雀において，強いコンピュータプレイヤは十分価値があると考える．

チェスなどと比較して麻雀が持つ複雑さの 1 つとして，麻雀ではプレイヤから見えている盤面の状態が同じでも，取るべき戦略が異なるという難しさが存在すると考えている．ここでいう麻雀の戦略とは，できるだけ早く上がる，より高い点数で上がるなどである．例えば，現在の順位が高い場合は，ゲームを早く終わらせるために（期待される点数が低くなってしまったとしても）できるだけ早く上がる行動をを取るという戦略が挙げられる．一方で，現在の順位が低く，上位の選手との点数差が大きい場合は，上がるのが難しい点数の高い手を作る戦略が有効となる場合が多い．また，そのどちらにも偏らない中間の戦略も考えられる．このように，麻雀には状況に応じて多彩な戦略が必要になること，また麻雀には利用されるルールが様々であり，それに応じて有効となる戦略が異なることが，難しさとして考えられる．これらは人の手によるコンピュータの戦略作成を困難にしている．そこで，これらの異なる戦略を機械学習によって自動的に得ることには価値があると考える．

先行研究では，実際の麻雀の牌譜を教師データとして評価関数を学習する手法 [5] や，モンテカルロ法の報酬を調整することで異なる戦略を得る手法 [6] が提案されており，人間の上級者程度の実力を獲得している．それらとは異なるアプローチとして，本研究では強化学習による麻雀プレイヤの作成を提案する．強化学習は，最適な方策をエージェントに自動的に学習させる機械学習であり，報酬や学習パラメータを変更するだけで，異なる方策を得ることが可能である．そこで，強化学習の報酬や設定を変えることで容易に多様な麻雀の戦略を得ることを目的とする．また，教師あり学習では多様なルールごとに教師データを収集し，それぞれのデータごとに学習を行う必要があるが，強化学習は教師データを収集する必要がないため，教師あり学習に比べて，多様なルールごとにそれぞれのルールに対応したプレイヤを作成することが容易であると考える．

初めに，麻雀を単純化したゲームに対して強化学習を適用し，実際の麻雀に適用する際に有効となりうる特徴量などの知見を収集する．麻雀を単純化したゲームとは，麻雀の要素を最低限残しつつ，そのルールを簡略化したものである．本研究では 5 種類のゲームを作成する．そこで得られた知見をもとに，実際の麻雀のルールに強化学習を適用する．

強化学習には（１）すべての状態とその状態で取りうる行動に対する状態行動価値をテーブルに保存し，それをエージェントの行動ごとに更新する“テーブル型”手法,（２）状態と行動を特徴量に変換し，その

特徴量に重みパラメータを掛け合わせたものの線形和を状態行動価値として，重みの更新により学習をする“特徴量型”手法,（3）ニューラルネットワークで状態行動価値を近似する“ニューラルネット型”手法の 3 つの手法を適用する．これらの手法には学習時間・学習精度・必要となるメモリの大きさなどに違いがあり，それぞれ長所と短所が異なる．

強化学習で異なる戦略を実現するために，報酬や学習パラメータを調整する．上がった時に一定の報酬を得る学習と上がった時の獲得点数を報酬とする学習を行う．前者は平均手数が少なくなるプレイヤ，後者は上がった時の平均点数が高くなるプレイヤを得ることを目的としている．さらに，これらの学習の割引率パラメータを調整した実験も行う．1 手ごとの期待報酬への割引が大きくなれば，多くの手数をかけて報酬の高い手を狙う戦略よりも少ない手数で低めの報酬を得られる手を選択する戦略が学習されると予想する．ほかにも，実験ごとに設定する行動回数の上限の違いによって同様の効果は得られると思われる．

本章に続き，第 2 章では本研究で対象とする麻雀におけるルールと重要な戦略などを述べる．また，本研究で扱う 1 人麻雀のルールもこの章で説明する．第 3 章では本研究に関連した，麻雀に関する学術的論文や，本論文で扱う多層ニューラルネットワークによる状態行動価値の近似を不完全情報ゲームに適用した論文を紹介する．第 4 章では本論文で扱う 3 種類の強化学習と，与える報酬によって得られると予想される戦略について述べる．第 5 章では本論文が提案する単純化ゲームついて説明し，それに強化学習を適用した場合の評価結果を述べる．その評価結果を受けて，1 人麻雀実験に継承すべき要素について考察する．第 6 章では本論文で扱う 1 人麻雀に強化学習を適用した実験と，その評価結果を述べる．本論文の主題である複数の戦略が得られていることを確認する．最後に第 7 章で本研究を総括し，今後の展望や課題を述べる．

第2章　対象ゲーム

本研究では，不完全情報ゲームである麻雀を対象とする．本章では，まず麻雀の基本的なルールについて概説する．次に麻雀における戦略を説明し，その多様性を述べる．最後に麻雀から多人数性を排除した1人麻雀のルールについて説明する．

2.1　麻雀のルール

麻雀は4人のプレイヤで点数を競い合う不完全情報ゲームである．各プレイヤにはゲーム開始時に13枚の手牌が配られ，順番に手番を行う．プレイヤは自分の手番ごとに山から1枚牌を引き，手牌から1枚牌を捨てるということを行いながら，14枚の牌の組み合わせからなる役の完成を目指す．役を完成させることを和了（ホーラ，日本語ではしばしばアガリとも）と呼び，1人のプレイヤが和了した場合，そのときの役に応じてほかのプレイヤから得点を得る．プレイヤのうちの1人が和了するか，山がなくなるまでを局と呼び，通常のゲームではこの局を特定の回数行い，すべての局が終了した時点での点数の高さを競う．牌は全部で34種類あり，それぞれの牌が4枚ずつあるため，ゲームでは合計136枚を用いる．牌には1から9までのいずれかの数字が掛かれた数牌と，文字が書かれた字牌が存在する．数牌には萬子（マンズ），筒子（ピンズ），索子（ソーズ）の3種類があり，合計で27種類存在する．字牌には東，南，西，北の4種類からなる風牌と，白，發，中の3種類からなる三元牌がある（図2.1）．和了に必要な基本的な条件は4つの面子と1つの雀頭を揃えることである．面子には，同じ種類の牌を3枚揃える刻子と同色の数牌で数字が順番に並んだ3つの牌の組み合わせの順子がある（図2.2）．雀頭は同じ種類の牌を2つ揃えたものを指す．

そのほかに本論文で扱う麻雀の専門用語を解説する

翻　全ての役ごとに割り当てられた数値．和了したときの得点の要素となる．和了した場合，手牌で構成することのできるすべての役の翻の合計で得点が決定される

聴牌　あと1枚特定の牌を引くことで和了することができるような状態

立直　聴牌したときにほかのプレイヤに自分が聴牌したことを知らせる行動．リーチをした後は和了するまで手牌の変更をすることはできない．1翻の役が付く

向聴数　聴牌になるのに必要な牌の枚数

有効牌　向聴数を少なくすることができる牌

放銃　自分の捨て牌でほかのプレイヤが和了すること．振り込みとも呼称する

降り　自分の和了をあきらめ，ほかのプレイヤに放銃しないような捨て牌の選択をすること

流局　誰も和了することなく山がなくなること

不聴罰符　流局時に聴牌していないプレイヤが聴牌しているプレイヤに払う点数

オーラス　ゲーム全体で最後の局のこと

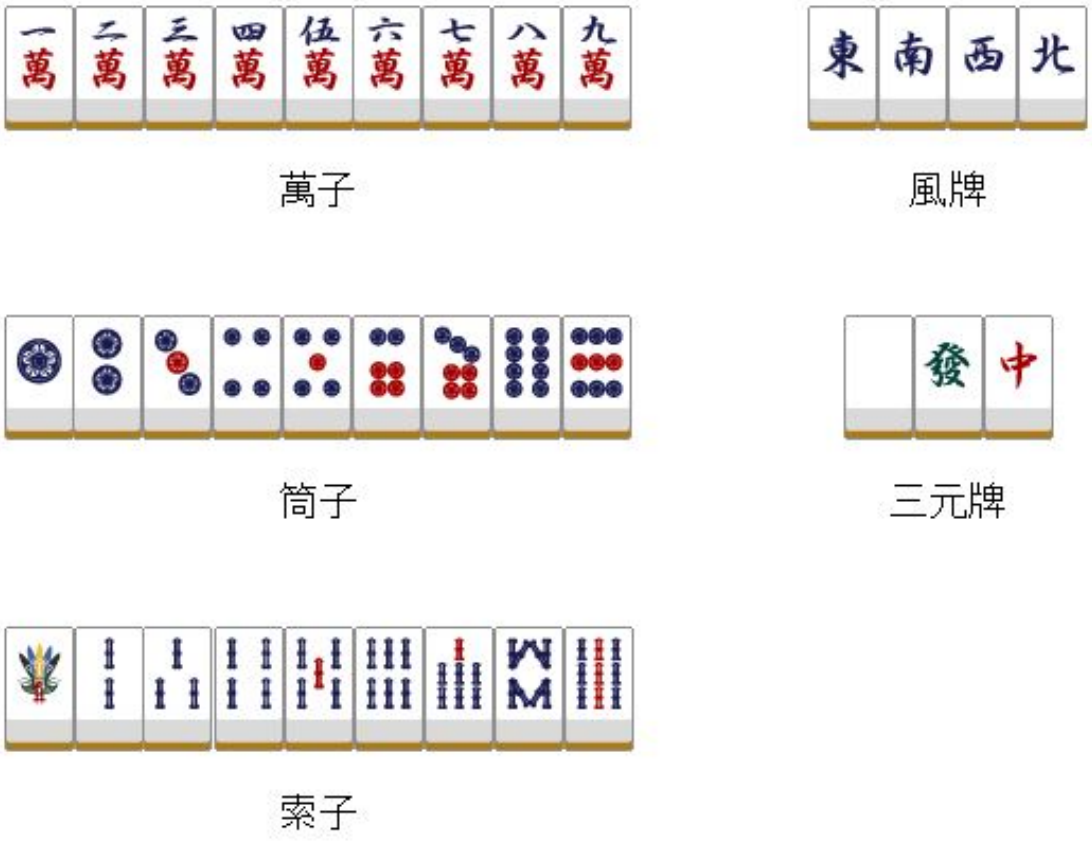

図 2.1: 麻雀で用いる牌

図 2.2: 面子の例

2.2　麻雀における戦略

麻雀の初心者の多くは手を決定するときに,「現在の順位状況や手牌にかかわらず和了することを目指す」ということを考えていると思われる．しかし，数局の最終的な得点を競う麻雀では，状況に応じて「点数の高さよりも手牌の完成の早さを急ぐ」,「得点を稼ぐために難しい役を和了する」「和了を目指すのに良くない手牌なので，和了をあきらめ放銃を防ぐ」など多様な戦略が考えられる．そのため，一般的な中級者以上の実力を持つ人間プレイヤは現在の自分の手牌や場の状況，点数と順位の状況を考えていると思われる．

また，麻雀において早く和了することと高い得点を和了することは基本的にトレードオフな関係にあるが,「それなりの早さでやや高い手を和了しよう」といった中間的な戦略も考えられる．例えば,「オーラスで自分が子で順位が 2 位，トップのプレイヤとの点差が 3000 点という状況で 3 位のプレイヤがリーチをかけた」ときに，自分の手が図 2.3 のような場面を考える．

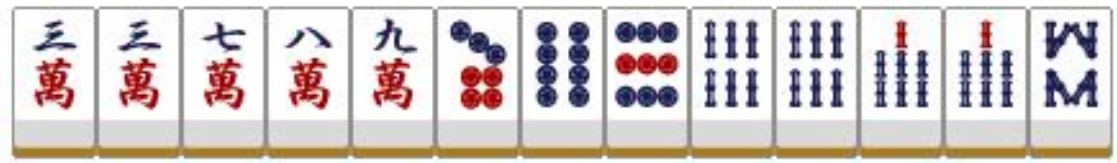

図 2.3: 手牌の例その 1

8 ソウを引けば 2000 点で和了となるが（図 2.4），この場合この手を和了しても順位の変動はない．

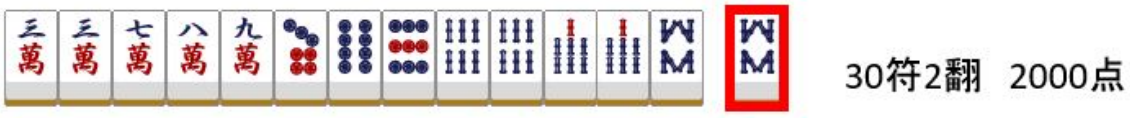

図 2.4: 2000 点で上がる戦略

9 ソウを引くのを待ち，5 ソウで和了すると 3900 点であり，順位を上げることができる（図 2.5）．

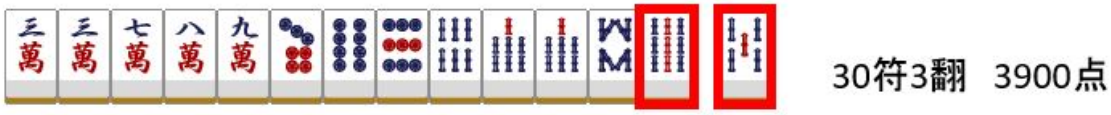

図 2.5: 3900 点で上がる戦略

また，さらに高い得点を狙うなら 8 ソウと 9 ソウを引くのを待ち，9 ソウで和了する手が考えられる（図 2.6）．

得点は 7700 点であり，この場合でも順位は上がる．しかし，ほかのプレイヤが上がる前に図 2.5 の場合よりも多くの牌を集めなければならない．ほかのプレイヤに和了され 2 位にとどまる，あるいは順位を落とす可能性が高いという点で図 2.5 の場合よりもリスクが大きい．それにもかかわらず，どちらも順位が 1 上がるという同じ結果となる．よって，この場合は図 2.5 の戦略を取ることが望ましい．

図 2.6: 7700 点で上がる戦略

また，次の図 2.7 について考える．このとき 9 ピンを捨てた場合は，4 ピン，6 ピン，7 ピンで和了することのできる聴牌となる．和了点数は図 2.8 の通りである．

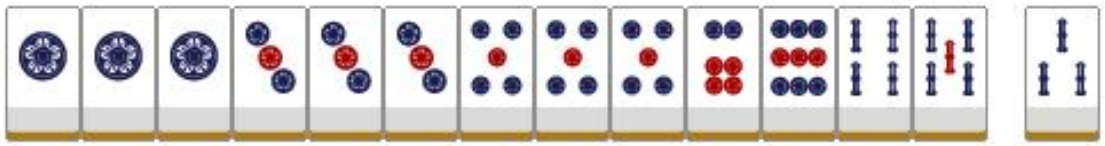

図 2.7: 手牌の例その 2

図 2.8: 多くの牌で待つ戦略の和了点数

一方で，3 ソウや 5 ソウを捨てる戦略について考える．この場合の向聴数は 2 となり，8 ピンを捨てる手よりも和了にかかる手数が多くなる可能性が高い．しかし，この場合であれば混一色や清一色といった翻数の高い役を狙うことができる（図 2.9 参照）．

そのほかにも，局の序盤で自分があと 10 回以上手番を行うことができる場合と，局の終盤で 3 回しか手番を行うことができない場合でも戦略は異なる．残りの手番が多い場合は必要な牌の種類が多くなったとしても集めることができる可能性が高いため，点数の高い手や待ちの広い手（有効牌の種類が多くなるような手）を作る戦略が有効である場合が多いと考える．一方，残りの手番が少ない場合はとりあえず聴牌を目指すことが和了や不聴罰符による失点を減らすことにつながる戦略になると考える．

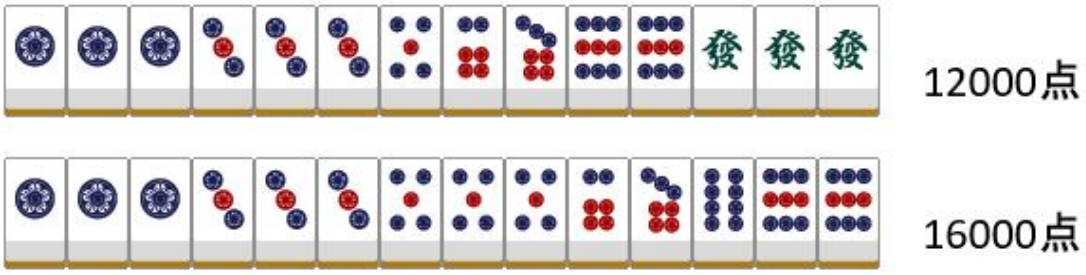

図 2.9: 高い点数を狙ったときの和了例

2.3　1 人麻雀

1 人麻雀は水上らの研究 [8] や，海津らの研究 [6] で用いられたゲームで，通常の麻雀から多人数性を排除したものである．多人数ゲームとしての複雑さはなくなるが，通常の麻雀における役を作り和了するという目的は残っているため，麻雀の部分問題の 1 つとして扱うことができる．以下に 1 人麻雀のルールを示す．

（１）山からランダムに 13 牌引き，初期手牌とする

（２）山から 1 枚牌を引く

（３）和了かどうかの判定を行う

（４）（３）で和了でなければ手牌から 1 枚牌を捨てる

（５）（４）が終了したら（１）に戻る

（６）（３）で和了であれば局を終了する

（２）から（５）までの流れを 1 回の行動とし，1 手と呼称する．本研究では，先行研究と同様に手の回数が一定数の場合にゲームを終了する方法と，先行研究では使われていない，和了となるまで無限に手を行う方法の 2 種類の方法を用いている．また，もう 1 つ先行研究の設定と異なる点として，山をカウントせず，どれだけ手が進んでも，すべての牌を等しい確率で引くことができることを挙げる．

これらの設定を用いる理由は 2 つある．1 つ目の理由は，学習序盤でも状態行動価値を得られるようにするという目的である．強化学習では，学習の序盤は指標となる状態行動価値を得られていないため，ランダムに行動することとなる．そのため，山の中身の制限や手数の制限を行うと，なかなか和了することができず，学習のための報酬を得るために膨大な時間が掛かることが予想される．

2 つ目の理由は，この単純化した設定のもとであれば強化学習における状態行動価値の理論値が得られるということである．十分に学習が進んだ強化学習では状態行動価値がこの理論値に収束するはずであり，それを確認するために単純化した設定を用いる．

この単純化設定の欠点として，和了に必要な牌を必ず引くことができるという前提で，コンピュータプレイヤが学習されてしまうことが挙げられる．この対策として，学習序盤は（学習促進のため）回数の制限を行わず，ある程度学習が進んだら（現実に近付け）手の回数に制限を設けることや，残り行動回数を状態に含めて価値関数を学習することなどが挙げられる．

もう一つの問題点として，残り牌枚数が少ない待ち方を忌避できないということが挙げられる．例えば，3,3,8,8 というシャボ待ちは，最大でも残り 4 枚しかない．これに比べ 4,5 などといった両面待ちは同じ 2 枚待ちでも最大 8 枚残っている可能性がある．このように本研究で用いている一人麻雀はいくつかの単純化を行っているが，将来的にはこれらも状態に含めることも可能であると考えている

第3章　関連研究

本章では，本研究に関連する既存研究を紹介する．

3.1　牌譜を用いた教師あり学習による1人麻雀プレイヤ

初期の水上らの研究 [8] では，まず 1 人麻雀プレイヤを教師あり学習を用いて作成し，そこに降りの判断を加えることで中級者レベルの 4 人麻雀プレイヤを作成することに成功した．

1 人麻雀プレイヤは，評価関数を教師あり学習を用いて作成している．教師データとしてオンライン麻雀サイト「天鳳」[9] の鳳凰卓の牌譜を用い，学習には牌譜との一致を目指した平均化パーセプトロンを使用している．ただし，牌譜をそのまま用いると 4 人麻雀特有と考えられる局面があるため，ある 1 局において初めてリーチをしたプレイヤがリーチをかけるまでの局面を教師データとしている．教師データの数は 170 万局面，特徴量は 37488 次元である．また，1 回のゲームでは 27 回ツモを行う．評価として 100 回のゲームでの和了率を人間上級者プレイヤ，人間平均プレイヤ，そして三木らの研究 [10] で用いられた PlainUCT と比較している．和了率は 48 %を示し，人間平均プレイヤよりも高く，上級者に近い実力を示した．

水上らはこの 1 人麻雀プレイヤに降りの判断を付与することで 4 人麻雀への拡張を図った．手牌にある各牌を，1 人麻雀プレイヤが切ろうとしている順番に順位付けし，その中の上位に実際の牌譜の捨て牌が存在しないとき，1 人麻雀と 4 人麻雀とで戦略に差が生じたとしている．1 人麻雀プレイヤが判断を誤った手について，人間上級者プレイヤが降り，や回し打ちなどのラベル付けを行ったところ，外した全局面のうち 42 %が降りに関するものであった．よって，水上らは 1 人麻雀に降りの判断を付与することで 4 人麻雀への拡張ができると考えた．牌譜の局面に降りるべき局面かどうかのタグ付けを手動で行い，そのデータを教師データとして，入力した局面で降りるべきかどうかを判断する評価関数を学習した．カッパ係数は 0.75 となり，概ね降りるべきかどうかの判断ができたとしている．

最後に降りの判断を付与した 1 人麻雀プレイヤをオンライン麻雀サイト「天鳳」で対戦させることで性能を評価している．レーティングは 1507 点で人間の平均プレイヤと同等の実力を示した．

3.2 複数の予測モデルを組み合わせた麻雀プレイヤ

水上らの研究 [5] では，麻雀が 4 回から 8 回の部分ゲームを行う繰り返しゲームであることと，部分ゲームの各収支ではなくすべての部分ゲームが終了した時点の点数状況が勝敗を決定することから，長期的な戦略が重要であることに着目した．そこで，モンテカルロ法を用いた水上らの研究 [14] のモンテカルロのシミュレーションの報酬を，1 局の収支ではなく，すべてのゲームが終了した時点の期待最終順位とするプレイヤを提案した．

また，局の序盤からモンテカルロ法を用いると，精度の高い手を選択することは困難であるため，序盤は先行研究 [8] で用いていた 1 人麻雀プレイヤで手の決定を行い，特定の条件を満たした場合にモンテカルロ法による手を用いるコンピュータプレイヤを利用した．条件は，ほかのプレイヤがリーチをかけたとき，ツモ可能な牌の枚数が 16 以下，すなわち局の後半になったときなどである．

順位の予測は，最終順位を予測する多クラスロジスティック回帰モデルを使用して，1 位から 4 位を予測する多クラス問題とした．出力にソフトマックス関数を用いて，現在の得点状況から期待最終順位を出力するモデルを作成した．学習には「天鳳」の牌譜を用いた．現在の順位をそのまま最終順位として返すベースラインと比較を行ったところ，実際の順位と予測の平均絶対誤差はベースラインが 0.809 に対し，得られたモデルは 0.763 となり，得られた予測モデルの方が精度が良かった．

最後に，このモデルを搭載したコンピュータプレイヤを，オンライン麻雀サイト「天鳳」で対戦させることで性能を評価している．安定レーティングは 1844 で，先行研究 [14] の 1718 を上回り，天鳳の特上卓でプレイできる程度の実力を示した．このことから，現在までに発表されているコンピュータ麻雀プレイヤの学術的研究の中でも特に完成度の高い研究であると思われる．

3.3　モンテカルロ法を応用した一人麻雀プレイヤ

海津らの研究 [6] では，簡単なパラメータ調整のみで麻雀の多様な戦略を得ることを目的として，モンテカルロ法を応用して 1 人麻雀プレイヤを構築している．はじめに，単純なモンテカルロ法を 1 人麻雀に適用すると，プレイアウト中の行動をランダムに行ってしまうためほとんどのプレイアウトで報酬を得ることができないため，1 人麻雀には有効ではないことを示した．

そこでこの研究では，各手牌ごとに，その手牌を捨てると「どのくらいの点数が取れそうか」「どのくらい早くあがれそうか」をモンテカルロシミュレーションによって推測する試みを行う．“ 現在の手牌＋（残り順目数の）ランダムなツモ ” を与え，その集合の中から獲得点数最大の 14 枚を抜き出す．この 14 枚に含まれていない牌は “ 不要な牌＝捨てたい牌 ” であり，現在の手牌からの交換枚数が少ないほど早くあがれる手ということになる．

実際，点数の重みを大きくした場合，和了時の最高平均点数は 100 回のゲームで 11311 点で，和了するまでの手数は 15.89 手で，人間上級者プレイヤ（5936 点，14.16 手）を獲得点数の面で大きく上回っているが，和了できる確率 9 %と，和了までの平均手数は人間初心者プレイヤ（20 %，14.50 手）を下回っている．

早上がりの報酬に重みを置いた場合，平均手数は 14.50 手，平均点数は 1255 点となり，平均点数は大きく下がったが，手数は人間初心者プレイヤと同じ程度の実力にはできた．

3.4 麻雀の複数の戦略に対する着手モデル

田中らの研究 [11] は，麻雀初心者への教育をする際には，とるべき戦略と現在の状況でその戦略を選択する理由を示すことが必要であると考え，状況に応じた戦略の選択を行うモデルを提案した．そこで，麻雀の局面の情報からとるべき戦略を出力する決定木の作成を行っている．はじめに「天鳳」の麻雀大会，天鳳名人戦で実際に選ばれた手，計 991 手について手動で戦略の推察を行う．それを 5 つの戦略の組み合わせに分類し，その中の「早い和了を目指す」,「振り込みを避ける」,「高得点を目指す」の 3 つの戦略についてそれぞれの単目的行動モデルを作成している．

「早い和了を目指す」戦略モデルは，牌を捨てた後の向聴数が，捨てる前の向聴数以下になるような捨て牌を選ぶ．候補となる捨て牌が複数ある場合は，捨てた後の有効牌の枚数が最も多い手を選択している．

「高得点を目指す」戦略モデルでは，麻雀の役の中で出現頻度が高く，かつほかの役と重複しやすいドラとタンヤオを狙おうとするモデルを作成している．

「振り込みを避ける」戦略モデルでは，初めに他のプレイヤの捨て牌，現在の順目，リーチの有無という入力から，手牌にあるすべての牌に対して，その牌を捨てたときの「安全さ」の導出を行う．この安全さは，予測される振り込みとならない確率である．上級者の牌譜から統計をとった順目における，各プレイヤの聴牌確率を用いて導出している．

「早い和了を目指す」,「高得点を目指す」のそれぞれのモデルについて一人麻雀で評価を行ったところ，前者のモデルは 100 回のゲームで人間平均プレイヤの和了率を超え，後者のモデルは前者のモデルに比べてドラとタンヤオの出現率が大幅に増加し，和了時の平均点数も 1.7 倍まで増加している．

「振り込みを避ける」モデルの評価では「天鳳」の牌譜から 22828 のツモ局面について全牌種についてほかのプレイヤに振り込む確率を予測させた．実際に 1 人以上のプレイヤがその牌で待っていた割合と比較したところ，おおむね実際の割合に近い値を予測できていた．

初心者教育を行うためには，これらのモデルが上級者の戦略を判断できているかが重要なため，上級者の手に対してこれらのモデルで分類を行う．3 つのモデルすべてで高く評価されたもの，2 つのモデルで高く評価されたもの，1 つのモデルのみで高く評価されたものという分類を行い 7 つのタイプに分けられた．この 7 つのタイプを予測する決定木を *J*4.8 アルゴリズムで学習し，考査検証法を用いて性能の評価をしたところ，正しく導き出せた割合は 32.8 %となった．

最後にこのモデルを用いて上級者の手を再現できるかを実験した．はじめに，与えた局面について決定木で 7 つのタイプのどのタイプに分類されるかを調べる．その後に実際の捨て牌がモデルが選択しようとする上位 3 位以内に入っているかを調べる．タイプが一致した場合は上位 3 位以内に入る確率が 100 %と非常に高い精度を示し，タイプが不一致でも 80.5 %の確率となり，全体の予測精度は 86 %であった．決定着の学習に用いた特徴量の吟味を行うことで精度の向上ができると考察している．

3.5 麻雀への多層ニューラルネットワークの適用

築地らの研究 [12] では，不完全情報ゲームにおいて強いコンピュータプレイヤを作成するためには複雑な評価関数が必要であることと，複雑な評価関数を作成することができる Deep Learning にかかわる技術が目覚ましく進歩したことから，不完全情報ゲームである麻雀に対して Deep Learning の適用を行っている．また，既存のゲーム研究の多くは，評価関数を作成する際にそれぞれのゲーム特有の特性を利用しているものが多いことに触れている．Deep Learning の特徴である，人間の知識を排除した設計でも評価関数を学習できるという点に着目し，この研究では面子や雀頭といった麻雀特有の特徴の組み合わせを，できるだけ特徴量として用いずに人間プレイヤの行動選択を模倣できるか確かめようとしている．

学習データとして東風雀の牌譜データを用い，局面データを入力，実際に捨てられた牌を教師として，ある局面において牌譜と同じ牌を捨て牌に選択できるようなネットワークの獲得を目指している．ソフトマックス関数で捨て牌を選択する設計として，麻雀の牌 34 種類のうち，どの牌を捨て牌として選択するのかという多クラス分類問題としている．特徴量の次元数は 1653 である．

ネットワークを複雑にすることで学習データに対する一致率は 75.1 %まで向上したが，未知のテストデータへの一致率は 40.8 %と汎化性能は低い．これについて，ネットワークを複雑にすることによってモデルの表現力を上げることに成功したが，勾配消失問題や過学習が起きたことが原因であると考察している．

萩原らの研究 [13] では，麻雀において不確定な情報を推定する精度が向上すればコンピュータプレイヤの実力の向上につながると考え，機械学習を用いて，局面の情報から相手プレイヤの和了点数の予測を行っている．はじめに，特徴量のグルーピングが和了点数の予測精度向上につながっていることを示し，さらに複雑なモデルで精度が向上するかを確かめるため，特徴量のグルーピングを非明示的に行っていると考えられる多層ニューラルネットワークを用いて学習を行った．

学習には「天鳳」の牌譜から 30 万局面を用いている．特徴量の次元数を 183 とし，予測される翻数とテストデータの翻数との平均二条誤差平方根で汎化性能を評価している．中間層 3 層，中間ノード数 300 で学習を行ったところ，汎化性能は 0.60090 となった．これは，水上らの研究 [14] で示された学習局面数 5920 万，特徴量の次元数 26889 として，教師データに対する重回帰モデルを用いて学習した場合の汎化性能 0.60828 を上回る結果である．萩原らは，水上らの研究で用いられた局面数はわずかに 100 局面であり，有意差があるかは明らかでないとしている．

3.6 不完全情報ゲームにおける多層ニューラルネットワークによる強化学習の価値関数の近似

佐藤らの研究 [15] では不完全情報ゲームである花札のこいこいの強いコンピュータプレイヤの作成を行った．花札は不完全情報ゲームであり，麻雀に形式が似たゲームであるが，上級者の棋譜を大量に用意しづらいという問題があり，佐藤らは寄付を用いた教師あり学習は困難だと考えた．そこで，強化学習を用いて花札のコンピュータプレイヤの作成を提案している．強化学習として方策勾配法と Neural Fitted Iteration [16] を適用している．Neural Fitted Iteration（以下 NFQ）は Martin らが提案した手法であり，ANN によって近似した価値関数を用いる点と，通常の Q 学習は 1 回の行動で環境から報酬を受け取るごとに価値の調整を行うのに対し，NFQ は事前に集められたデータの塊で学習を行い，価値を調整するという点の 2 点が特徴である．

実験では各カードがプレイヤから見てどの場所にあるのか，現在が何ターン目であるのか，相手の手札になさそうな月を特徴量としており次元数は 268 である．学習中の手の決定には $\varepsilon - greedy$ 法を用いている．訓練データとして入力を局面の特徴ベクトルとして，出力をその局面から「相手プレイヤの手札」と「山札」の中身とカードの順序をランダムにシャッフルしながらルールベースのプレイヤとの 100 回の対戦シミュレーションを行ったときの平均スコアを用いる．データが 60000 件に達するたびにニューラルネットワークの重みの更新を行う．更新ごとに，その時点のネットワークを用いて 1000 回対戦し性能を評価している．

実験結果は，ルールベースプレイヤ同士の対戦結果で先手の平均獲得スコアが+0.051 点なのに対して，得られたコンピュータプレイヤとルールベースプレイヤに+0.5 点程度と有意な差を得ている．

佐藤らの研究は麻雀に関するものではないが，問題の背景や用いた手法が本研究と似ており，直接参考にしたものでもあるため紹介する．

第4章　強化学習

この章では，本研究で用いる 3 つ種類の強化学習について説明する．また，強化学習のパラメータや報酬によってどのような戦略が得られるのかという予想について考察する．

強化学習 [7] はエージェントがある状態のときにとった行動の価値をもとに，最も多くの報酬が得られるような方策を学習する機械学習である．教師あり学習とは異なり，状態行動に対する明確な教師データは存在しない．その代わりに，エージェントは行動をとった際に環境から得る報酬を頼りに学習を進める．行動をとり報酬を得るということを繰り返すことで，最も多くの報酬が得られるような方策を学習する．

4.1　テーブル型

“テーブル型”では，全状態と，その時に取りうる全行動の価値をテーブルに保存する．エージェントが行動をとった際，テーブル内の価値を更新する．価値の更新には以下の更新式 4.1 を使用する．

$$Q(s,a) = Q(s,a) + \alpha \left[R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \tag{4.1}$$

状態 s のときに行動 a をとったときの状態行動価値を $Q(s,a)$ とする．α は学習率であり，γ は将来の報酬に対する割引率を示す．$R(s,a)$ はその時に環境から与えられる報酬を表し，$\max_{a'} Q(s',a')$ は s の次状態 s' における状態行動価値の最大値を表す．よって，この更新式はエージェントが状態 s で行動 a をとった時の価値を，次の状態で最も高い値を持つ状態行動価値に, 将来の報酬に対する割引を考慮した値へと近づけるという学習を表している．

本研究では，状態を手牌，行動を捨て牌とする．牌を捨てるたびに，山から 1 枚牌を引く．牌を引いたあとにアガリかどうかのチェックを行い報酬を与える．ここまでを 1 つのエピソードと呼称する．和了した場合はそのエピソードを終了し，次のエピソードを始める．テーブル型の最大の特徴は，学習率を適切にスケジューリングした場合，すべての状態行動価値が正しい値に収束することが保証されている点である．行動をすべてランダムで行った場合でも，最終的に状態行動価値が収束することは知られているが，一般的には収束を早めるために一定の確率 ε でランダムに行動し，それ以外ではテーブルを参照して，価値が最も高い行動をとるという方法（$\varepsilon - greedy$ 法）を用いる．

一方で，テーブル型の欠点として状態空間が膨大になると，すべての状態行動価値をメモリに保存することができなくなるという欠点がある．また，ほとんど同じ状態であっても，違う状態として別々に学習するため，汎化が生じず学習に長い時間が掛かる可能性がある．これらの点を踏まえて，本研究では後述する 2 つの手法も用いることとする．

4.2 特徴量型

テーブル型はすべての状態ごとにそれぞれの価値をテーブルに保存しなければならない．そのため，状態空間が膨大になると，必要なメモリが大きくなるというデメリットがある．そこで，現在の状態を特徴量で表し，その特徴量ベクトルと同じ次元数を持つ重みベクトルの積で，状態行動価値を表現する手法を提案し，本研究では“テーブル型”と呼称する．

状態 s と行動 a を特徴量へと変換する式 $f_n(s)$，$g_k(a)$ を用いて，式 4.2，式 4.2 のように特徴量ベクトルを生成する．

$$s \to [f_1(s), f_2(s), f_2(s), \ldots, f_n(s)] \quad a \to [g_1(a), g_2(a), f_c(a), \ldots, g_k(a)] \tag{4.2}$$

重みベクトル $\vec{w}$ を定義して式 4.3 のように状態行動価値を表す．

$$Q(s,a) = \sum_{i=1}^{n}(w_i f_i(s)) + \sum_{j=1}^{k}(w_{(n+j)} g_j(a)) \tag{4.3}$$

行動を行うごとに重みベクトルを更新することで学習を行う．重みの更新式は式 4.4 となる．

$$\vec{w} += \alpha \left[R + \gamma \max_{a'} Q_w(\vec{w}, s', a') - Q_w(\vec{w}, s, a) \right] \cdot \nabla_{\vec{w}} Q(\vec{w}, s, a) \tag{4.4}$$

特徴量型は，テーブル型に比べて必要なメモリが少なくなること，類似する状態の価値を一度に更新することができるために，収束が早まるという点で優れるが，パラメータ数やモデルによっては最適解に対する表現力が不十分であり，テーブル型に比べて性能が劣る可能性があるというデメリットを持つ．

なお，本研究で扱う一人麻雀では，例えば「1,1,4,5,8 で 8 を捨てる行動」と「1,1,2,4,5 で 2 を捨てる行動」は，以降の期待報酬という意味で完全に同一である．従って，“現在の状態と行動の価値”を学習する（テーブルや特徴量で表す）のではなく，“捨てたあとの状態の価値”を学習する（表す）ことにする．

4.3 ニューラルネット型

特徴量型で十分な表現力が得られないことがある理由の 1 つは，状態行動価値を表現するにはモデルが単純すぎて異なる状態行動の価値を異なるものとして表現できないからだと考えられる．特徴量型では，線形モデルを用いて価値を表現していた．そこで，より複雑なモデルとしてニューラルネットワークで価値を近似することを提案する．本研究ではこれを“ニューラルネット型”と呼称する．佐藤らの研究 [15] で用いられた NFQ を用いる．学習には手牌の特徴量ベクトルとその手牌の割引報酬を用いる．特徴量型と異なる点として，状態行動価値を保存するのではなく，ニューラルネットワークによって予測される価値を用いることが挙げられる．ニューラルネット型では

1. ε-greedy 法に基づき 1 ゲームプレイする．その系列を捨てた後の手牌の列 $s_1, s_2, ..., s_k$ を保存する．
2. 最終的に和了できなかった場合は，今回の価値として 0 をペアとして与え $(s_1, 0), (s_2, 0), ...$ とする．
3. 最終的に和了できた場合は，固定値または得点に基づく報酬 r をペアとして与え，割引率 γ を考慮して，

   $(s_k, r), (s_{k-1}, r\gamma), (s_{k-2}, r^2\gamma), ...$ とする．
4. このように（捨て牌後の）状態 s_i と今回の価値 q_i を一定ゲーム数求めたら，

   ニューラルネットワークによる推定価値 $Q_{NN}(s_i)$ を，q_i に，学習率 α 分近づける更新を行う．

   $Q_{NN}(s_i) \Leftarrow (1-\alpha)Q_{NN}(s_i) + \alpha q_i$
5. 政策が更新されたので，終了条件を満たすまで，1. に戻り繰り返す．

4.4　学習パラメータと得られる戦略

4.4.1　割引率

割引率は，将来得られる報酬に対する割引であり，これは「将来のことはあてにならないため，多少報酬の値が低かったとしても，近く得られる利益を優先する」ような戦略を表現したパラメータであり，早くゴールに近づくことを目標とするような強化学習ではほぼ必ず使われる．

図 4.1 を用いて割引率が戦略決定にかかわる場面の例を解説する．現在の状態から 1 回行動すると 70 点の報酬を得る行動と，3 回行動すると 100 点の報酬を得る行動がある．

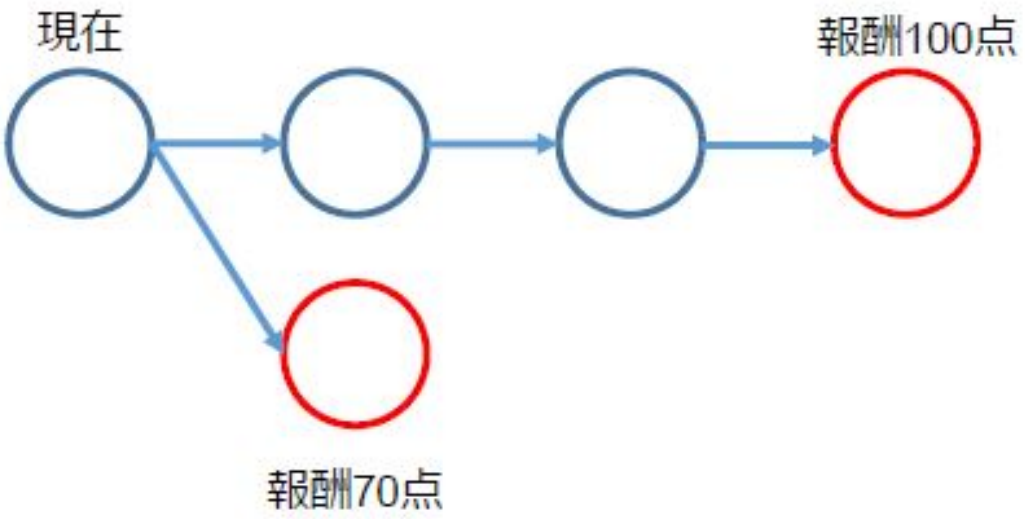

図 4.1: 状態遷移の例

割引率を 0.9 とすると

1. 上の行動の割引報酬　$100 \times 0.9^3 = 72.9$
2. 下の行動の割引報酬　$70 \times 0.9 = 63.0$

となり，上の行動の方が状態行動価値が高くなるため、上の行動が選ばれる．
次に，1 手ごとの割引を大きくした場合を考える．割引率を 0.7 とすると

1. 上の行動の割引報酬　$100 \times 0.7^3 = 34.3$
2. 下の行動の割引報酬　$70 \times 0.7 = 49.0$

となり，下の行動の方が状態行動価値が高くなるため、下の行動が選ばれる．
このように，割引の大きさが大きくなると，行動回数の少ない方策が得られやすくなる．

麻雀では，比較的和了しやすい低い点数の手と，和了しにくい（あるいは完成までに手数がかかる）高い点数の手というのは本質的にトレードオフの関係にあり，これのバランスをとるためのパラメータとして割引率を求めることができる．

4.4.2 報酬

本研究では，強化学習の報酬として以下の 2 種類の報酬を設定している．

(a) 和了時にできた役にかかわらず一定の報酬を与える

(b) 和了時にできた役に応じた獲得点数を報酬として与える

(a) の設定では，和了までの手数が少なくなるような方策が学習される傾向がある．どのような手で和了しても報酬の値が同じなので，和了までの割引回数が少ない行動が選ばれやすくなるからである．

一方で，(b) の設定では，和了時の点数が高くなるような方策が学習される傾向がある．

第5章　単純化ゲーム

本章では，本研究で設定した単純化ゲームについての説明と，それらのゲームに強化学習を適用した実験について述べる．また，本研究では 5 つの単純化ゲームに対して 3 つの手法を適用しているため，初めにどのゲームにどの手法を適用したのかを表 5.1 に整理する．表 5.1 の○が本研究で設計した実験である．×は必要なメモリが大きすぎるなどの理由で設計できなかった問題である．—はゲームに対してモデルが高級すぎるため省略した問題である．

表 5.1: 本研究で扱う単純化ゲーム

単純化ゲーム	手牌の数	牌の種類	状態数	テーブル型	特徴量型	ニューラルネット型
3 枚麻雀	3	4	20	○	○	—
1 色 5 枚麻雀	5	9	1287	○	○	—
2 色 5 枚麻雀	5	18	26334	○	○	—
2 色 8 枚麻雀	8	18	1.1×10^6	×	○	—
3 色 8 枚麻雀	8	27	1.7×10^7	×	○	○
1 人麻雀	14	34	3.4×10^{11}	×	○	○

5.1 3枚麻雀

5.1.1 3枚麻雀のルール

3枚麻雀は，手牌の枚数を3枚として，麻雀における面子を完成させた場合を和了とする．牌の種類は0から3の4種類とする．はじめに，初期手牌としてランダムに2牌を引き，そのあとはランダムに1枚引き1枚捨てるということを繰り返し和了した場合，ゲームを終了する．

5.1.2 3枚麻雀のテーブル型実験

3枚麻雀のテーブル型の実験を行った．
学習パラメータは以下の通りである．

- 割引率 $\gamma = 0.9$
- 報酬 和了したときに100を与え，それ以外では報酬は与えない
- 総エピソード数 3.0×10^6
- 学習率 $\alpha = 0.3 \times \frac{10^5}{10^5+N}$ ，N は現在のエピソード数
  $(\alpha : 0.3 \to 0.01)$
- $\varepsilon = 0.3$

$\varepsilon - greedy$ 法を用いた強化学習では，学習後半で状態行動価値がある程度学習され，有効な行動選択が行えるようになった場合も一定の確率でランダムに動くため，学習率が高いままであると，最適さ状態行動価値に収束しない可能性がある．そこで，学習率は，学習が進むにつれて値が小さくなるように調整した．

表5.2にテーブル型の結果を示す．表5.2はそれぞれの手牌と捨て牌について，学習終了時に得られた状態行動価値を示している．

表5.2: テーブル型3枚麻雀の結果

手牌	捨て牌:0	捨て牌:1	捨て牌:2	捨て牌:3
0,0,1	81.4	78.6	-	-
0,0,2	80.3	-	78.4	-
0,0,3	70.8	-	-	78.6
0,1,1	83.5	79.8	-	-
0,1,3	79.0	70.7	-	80.8
0,2,2	83.6	-	81.1	-
0,2,3	81.2	-	70.8	78.7
0,3,3	78.1	-	-	70.7
1,1,2	-	90.9	83.2	-
1,2,2	-	83.1	91.1	-
2,2,3	-	-	78.5	83.8
2,3,3	-	-	77.3	79.8

学習が正しく進んだことを確かめるために，実験で得られた状態行動価値と理論値の比較を行う．3 枚麻雀の理論値は以下のようにして計算することができる．

- 牌を捨てた後，次の状態に遷移する確率を求める（すべて $\frac{1}{4}$ の確率）
- それぞれの次状態において，もっとも割引報酬が高くなるような行動 a' を求める
- 次状態 s' に遷移する確率と s' で a' を選んだ場合の割引報酬を掛け合わせる
- すべての状態の報酬を足し合わせる．

式に表すと式 5.1 となる．

$$Q_{(s,a)} = \sum_{a=0}^{3} \{ \frac{1}{4} (Reward_{s'} + \gamma \max Q_{(s',a')}) \} \tag{5.1}$$

例えば，手牌 (1, 2, 2) のとき 2 の牌を捨てる場合，次の状態は (1, 2) である．この問題では (1,2) で 2 枚待ちにすることは明らかに最善の待ち方である．(確率 $\frac{2}{4}$)

このとき，

- 0 か 3 を引けば和了で報酬を得る（確率 $\frac{2}{4}$）
- 1 か 2 を引いた場合，その牌を捨てて (1, 2) の状態に戻るのが最善の行動である（確率 $\frac{2}{4}$）

したがって，理論値は以下のように計算できる．

$$\begin{aligned} Q_{(1,2)} &= \frac{2}{4} \times Reward + \frac{1}{2} \times Q_{(1,2)} \times \gamma \\ Q_{(1,2)} &= \frac{2}{4} \times 100 + \frac{1}{2} \times Q_{(1,2)} \times 0.9 \\ Q_{(1,2)} &= 90.9 \end{aligned} \tag{5.2}$$

実際の学習では (1,2,2) で 2 を捨てる状態行動価値は 91.1 と計算されているため，ほぼ理論値と同じである．なお，3 枚麻雀実験でのみ (1,1,2) で 1 を捨てることと，(1,2,2) で 2 を捨てることは同一視していない．以降の実験ではテーブル型でも特徴量型でも次状態のみを考慮し (1,1,2) で 1 を捨てることと (1,2,2) で 2 を捨てることは同一視している．

実際に学習した状態行動価値が，理論値にどれほど近くなったのかは，学習の成功を判断する最もよい基準である．これを，各状態行動価値それぞれについて見るのではなく，テーブル全体を見て，理論値との誤差を評価するために，平均二乗誤差を求めることにした．また，総エピソードによって誤差に差が出るのかを調べるため，総エピソード数を 10^7 から 5.0×10^3 まで変化させて実験を行った．結果を図 5.1 にまとめる．

図 5.1 の結果から，エピソード数が多いほど理論値に近い値を得ることができるようになっていることがわかる．よって，エピソードを繰り返すたびに理論値に近い値へと学習が進んでいるといえる．このことから，テーブル型の学習で 3 枚麻雀は学習できたと考えている．

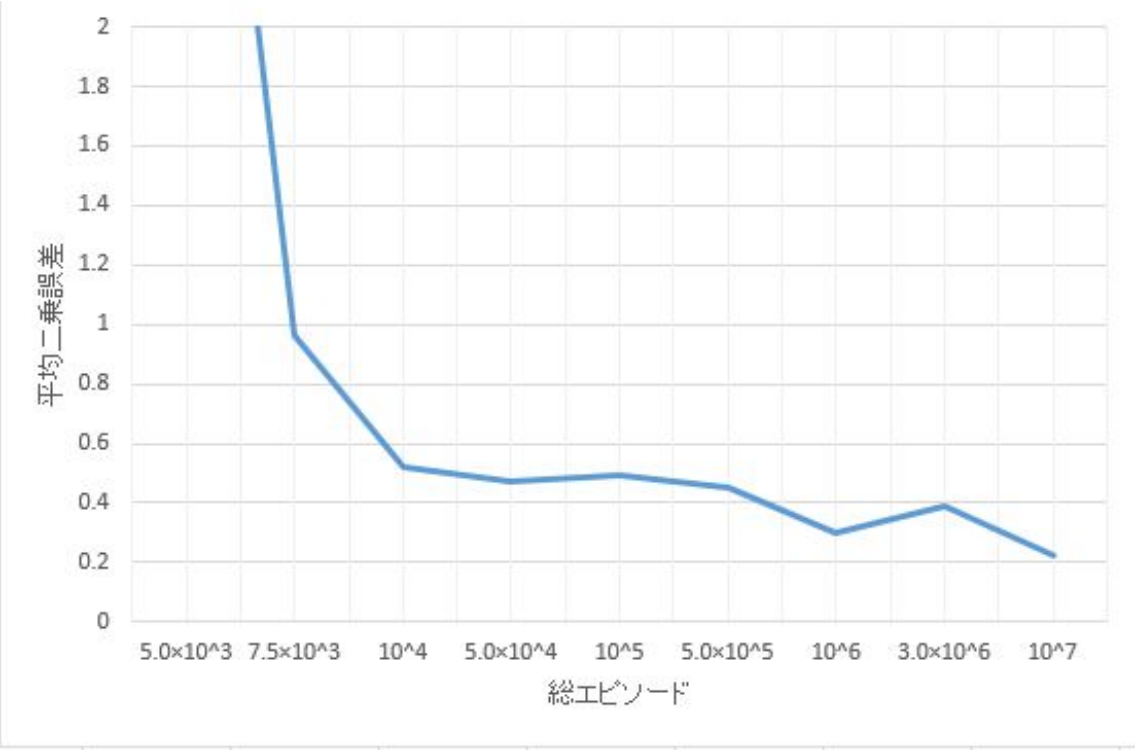

図 5.1: テーブル型 3 枚麻雀の理論値との比較

5.1.3 3 枚麻雀の特徴量型実験

特徴量型の手法が本研究の問題に適しているかを調べるため，3 枚麻雀の特徴量型の実験を行った．
行動の特徴量として，以下の 4 種類の特徴量を定義した．

- (0,0) または (3,3) の組み合わせを作れるか
- (1,1) または (2,2) の組み合わせを作れるか
- (1,2) の組み合わせを作れるか
- (0,3) の組み合わせを作れるか

また，状態の特徴量として，以下の 6 種類の特徴量を定義した．

- 0or3 の有無
- 1or2 の有無
- (0,1) の組み合わせの有無
- (1,2) の組み合わせの有無
- (2,3) の組み合わせの有無
- 同じ 2 牌の組み合わせの有無

学習パラメータは以下の通りである．

- 割引率 $\gamma = 0.9$
- 報酬 和了したときに 100 を与え，それ以外では報酬は与えない
- 総エピソード数 10^8
- 学習率 $\alpha = 0.3 \times \frac{10^7}{10^7+N}$，N は現在のエピソード数
  ($\alpha : 0.3 \to 0.03$)
- $\varepsilon = 0.3$

テーブル型の実験でエピソード数を増やすほど精度が高くなったことを受けて，テーブル型の最初の実験よりも総エピソード数を増やして実験を行う．それに伴い学習率の減衰速度を変更した．

また，同じ学習パラメータでテーブル型の実験を行った場合との比較を行う．比較方法として学習終了時に得られている状態行動価値で $\varepsilon = 0$（ランダムな行動は行わず，最も多くの報酬を得られるように行動する）のゲームを 10^6 回行い，その平均手数を比較する．学習曲線を図 5.2 に示す．3 枚麻雀ではテーブル型と特徴量型の手数の収束にかかるエピソードは，どちらの場合もほとんど変わらなかった．

最終的に得られた結果を表 5.3 にまとめる．

表 5.3 からどちらも同じ平均手数になるまで学習が進んだことがわかる．平均手数 3.2 については，聴牌の状態のときに和了牌を引く確率が $\frac{1}{2}$ から $\frac{1}{4}$ 程度であることを考えると，十分な値であると考えられる．また，平均二乗誤差もほとんど同じ値となった．特徴量型はメモリを削減できたり，収束を早くすることができる一方で，個別の状態行動価値を正しく表現できないという欠点を持つ．3 枚麻雀の問題では，状態数が非常に少ないため，テーブル型のデメリット（メモリ・収束速度の遅さ）は問題にならなかったが，一方

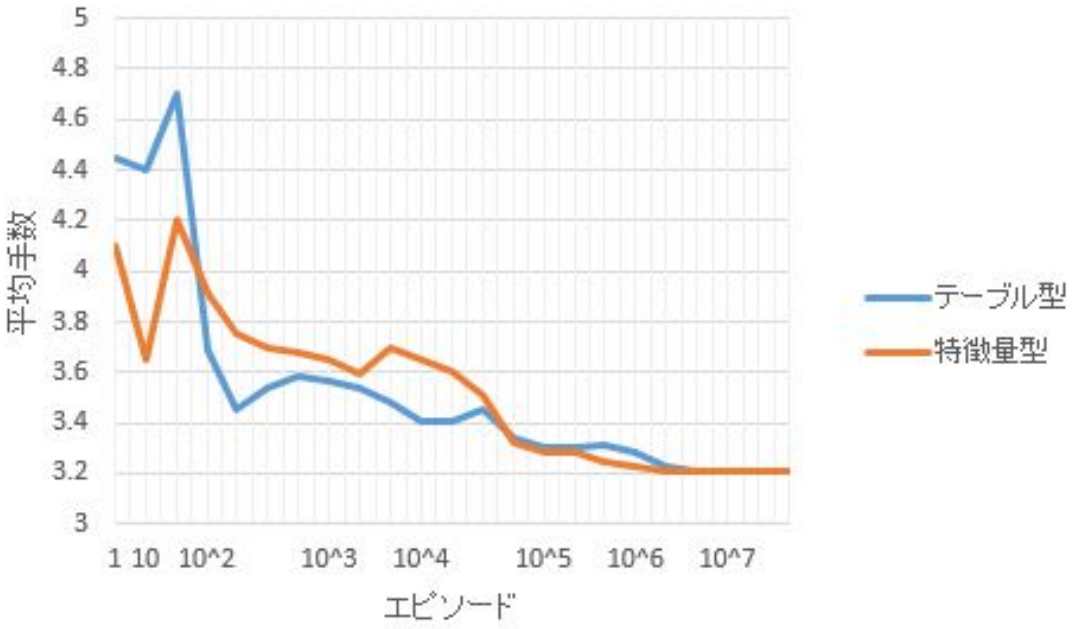

図 5.2: テーブル型と特徴量型の学習時曲線

表 5.3: テーブル型と特徴量型の比較

手法	和了までの平均手数	平均二乗誤差	実験時間（秒）
特徴量型	3.2	0.26	289.9
テーブル型	3.2	0.22	92.7

で特徴量型のデメリット（適切な状態行動価値にならない）も問題にならなかった．ただし，計算時間は多くなった．この結果から，学習をするのに十分な特徴量と学習モデルを設計することができたといえる．この結果から，3 枚麻雀に対して強化学習を適用することで，和了を目指すためのプレイヤを学習することに成功したと考える．

5.2　1 色 5 枚麻雀

3 枚麻雀の学習に成功したことを受けて，それよりも状態空間の大きい問題を設定して実験を行う．状態空間を大きくする方法は

1. 手牌の数を増やす
2. 牌の種類を増やす

が挙げられる．そこで，1 色 5 枚麻雀を設計した．

5.2.1　1 色 5 枚麻雀のルール

手牌の数は 5 枚，牌の種類は 0 から 8 までの 9 種類とする．和了の条件は雀頭 1 組と面子 1 組を両方とも揃えることである．3 枚麻雀（牌種類 4）では状態数が 20 だったが，1 色 5 枚麻雀（牌種類 9）では状態数 1287 と 60 倍以上に増えている．状態数が増えることにより，テーブル型の収束速度が遅くなることや，特徴量型についてはパラメータ数に比べて，表現したい状態行動価値が増えることで最適な値にならないことがデメリットとして挙げられる．これらのデメリットについて調べるため，1 色 5 枚麻雀についてテーブル型，特徴量型の実験を行う．

5.2.2　1 色 5 枚麻雀実験

テーブル型と特徴量型で実験を行った．特徴量型で用いた特徴量は 42 種類である．順子の有無や 1 つ差の数字の組み合わせなど 1 人麻雀で用いた場合でも有効と思われるようなものを設定した．また，先述の通りこれ以降の実験では次状態の手牌を特徴量で表現している．1 色 5 枚麻雀実験で用いる特徴量を表 5.4 に示す．学習パラメータは以下の通りである．牌の種類が増えたことなどから，学習後半にランダムに行動したときに与える影響が 3 枚麻雀に比べて大きいと考え，学習率が減少する速度を速くしている．

- 学習率 α: $0.3 \times \frac{10^4}{10^4+N}$ ，N は現在のエピソード数
  ($\alpha : 0.3 \to 3.0 \times 10^{-5}$)
- 割引率 $\gamma = 0.9$
- 報酬 和了したときに 100 を与え，それ以外では報酬は与えない
- 総エピソード数 10^8
- $\varepsilon = 0.3$

また，一定のエピソードが終了するたびに，その時点で得られている状態行動価値を用いて 10^6 回のテストを行い性能を比較した．

表 5.4: 1 色 5 枚麻雀の特徴量

特徴量	種類数
各牌の有無	9
（刻子でない）対子の有無	1
各牌の対子の有無	9
刻子の有無	1
各牌の刻子の有無	9
0or8 が 2 枚以上ある	1
0or8 が 1 枚ある	1
1or7 が 2 枚以上ある	1
1or7 が 1 枚ある	1
2or6 が 2 枚以上ある	1
2or6 が 1 枚ある	1
3,4,5 が 2 枚以上ある	1
3,4,5 が 1 枚ある	1
1 つ差の数字の組が 2 つ以上ある	1
1 つ差の数字の組が 1 つある	1
2 つ差の数字の組が 2 つ以上ある	1
2 つ差の数字の組が 1 つある	1

結果は図 5.3，表 5.5 の通りである．図 5.3 から 10^8 エピソードが終了した時点の平均手数はどちらも約 5.3 手とほとんど同じになっていることがわかる．聴牌のときに 2 種類の牌を和了牌として待っている場合は $\frac{1}{4}$ 程度の確率で上がることができるため，十分な値であると考えられる．このことから，学習をするのに十分な特徴量と学習モデルを設計することができたといえる．この結果から，1 色 5 枚麻雀に対して強化学習を適用することで，和了を目指すためのプレイヤを学習することにおそらく成功していると考える．

実験で理論値が得られているかを調べた．手牌 (2, 3, 4, 5) について考える．1 色 5 枚麻雀では 2 種類の牌をアガリ牌として待つのが最善な戦略であることが明らかである．手牌 (2, 3, 4, 5) の場合は，

- 2 か 5 を引けば和了で報酬を得る（確率 $\frac{2}{9}$）
- それ以外の牌を引いた場合，その牌を捨てて (2, 3, 4, 5) の状態に戻るのが最善の行動である（確率 $\frac{7}{9}$）

したがって，理論値は以下のように計算できる．

$$
\begin{aligned}
Q_{(2,3,4,5)} &= \frac{2}{9} \times Reward + \frac{7}{9} \times Q_{(2,3,4,5)} \times \gamma \qquad (5.3)\\
Q_{(2,3,4,5)} &= \frac{2}{9} \times 100 + \frac{7}{9} \times Q_{(2,3,4,5)} \times 0.9\\
Q_{(2,3,4,5)} &= 74.1
\end{aligned}
$$

テーブル型の実験では手牌 (2, 3, 4, 5) の実験値は 73.8 であった．理論値に近い値になっていることがわかる．一方で，特徴量型では，実験値は 66.3 と理論値に近い値を得ることはできていない．しかし，平均手数はテーブル型とほとんど同じである．これは手牌のすべての牌に対する状態行動価値の比較は間違っていないからである．例えば，手牌 (2, 3, 4, 5, 8) で 4 を捨てて (2, 3, 5, 8) と待つ形の状態行動価値の実験値は 43.1 と手牌 (2, 3, 4, 5) の実験値 66.3 よりも低い．この場合は，(2, 3, 4, 5) を待つ手が選択される．

3 枚麻雀に比べてテーブル型では 1 エピソード当たりの実験時間が約 4 倍長くなった．全体で手数が 1.7 倍多くなっていることや，捨て牌を決定する際に 3 枚麻雀では手牌の中の 3 つの牌についてその牌を捨てたときの状態行動価値を求めるのに対して，5 枚麻雀では 5 枚の牌に対して計算することなどを踏まえると妥当な値だと考える．また，特徴量型では実験時間は約 10 倍と大幅に増加している．特徴量の数が増えたことが原因であると考える．また，3 色麻雀に比べて，特徴量型では手数の収束が早くなっていた．3 色麻雀に比べ，麻雀特有の特徴を追加したことが原因であると考える．このことより，これらの特徴量を今後の実験でも使用することとする．

表 5.5: テーブル型と特徴量型の比較

手法	和了までの平均手数	実験時間（秒）
テーブル型	5.3	379
特徴量型	5.3	3132

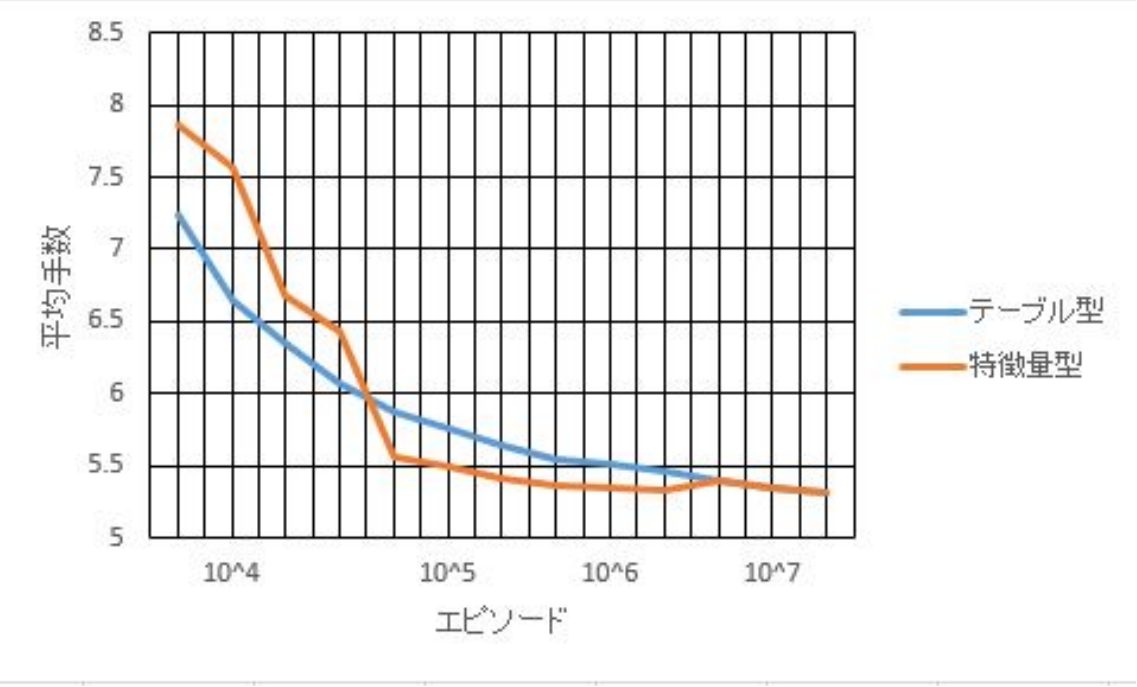

図 5.3: 1 色 5 枚麻雀の平均手数の推移

5.3　2 色 5 枚麻雀

3 枚麻雀から 1 色 5 枚麻雀へと手牌を増やした場合の学習に成功していることから，さらに状態空間の大きい実験を行う．実際の麻雀では 3 種類の数牌があることを踏まえて，複数の種類の数牌について実験を行い，学習が行われるかを試験したい．そこで，2 色 5 枚麻雀を設計した．

5.3.1　2 色 5 枚麻雀のルール

和了の条件は 1 色 5 枚麻雀と同じく，雀頭 1 組と面子 1 組を両方とも揃えることである．使用する牌は筒子と萬子である．通常の麻雀と同様に筒子と萬子を組み合わせて順子や刻子を作ることはできない．

5.3.2　2 色 5 枚麻雀実験

テーブル型と特徴量型で実験を行った．特徴量型で用いた特徴量は 76 種類とした．

また，学習率によって手数の収束がどのように変化するのかを調べるため，テーブル型については，複数の学習率を用いて実験を行った．

設定した学習率は，以下の通りとする．

実験 A　$\alpha = 0.3 \times (10^5/10^5 + N)$　$\alpha : 0.3 \to 7.5 \times 10^{-5}$

実験 B　$\alpha = 0.3 \times (10^6/10^6 + N)$　$\alpha : 0.3 \to 7.5 \times 10^{-4}$

実験 C　$\alpha = 0.3 \times (2.0 \times 10^6/2.0 \times 10^5 + N)$　$\alpha : 0.3 \to 1.5 \times 10^{-4}$

特徴量型　$\alpha = 0.3 \times (10^6/10^6 + N)$　$\alpha : 0.3 \to 7.5 \times 10^{-4}$

そのほかの学習パラメータは以下の通りである．

- 割引率 $\gamma = 0.9$
- 報酬 和了したときに 100 を与え，それ以外では報酬は与えない
- 総エピソード数 4.0×10^7
- $\varepsilon = 0.3$

一定のエピソードが終了するたびに，その時点で得られている状態行動価値を用いて 10^6 回のテストを行い性能を比較した．結果は図 5.4 の通りである．また，実験の後半部分を拡大したものを図 5.5 に示す．最終的に実験で得られた平均手数と実験にかかった時間を表 5.6 に示す．

表 5.6 から，実験 B と実験 C についてはすべてのエピソードが終了したときには，特徴量型よりも平均手数の面で，性能が高くなっていることがわかる．一方で，実験 A については特徴量型よりも性能が悪い．テーブル型は学習率のスケジューリングが適切に行われていれば状態行動価値が正しい値に収束するはずであるため，特徴量型よりも性能が良くなるはずである．よって，実験 B と実験 C については正しくスケジューリングされたが，実験 A はできていないことが考えられる．この結果から，2 色 5 枚麻雀に対して強化学習を適用することで，和了を目指すためのプレイヤを学習することにおそらく成功しており，学習率のスケジューリングによって性能が変化するという知見を得られたと考える．1 エピソード当たりの実験時間はテーブル型と特徴量型ともに約 1.5 倍に増えている．手数の増加や状態空間が大きくなったことが原因と考える．

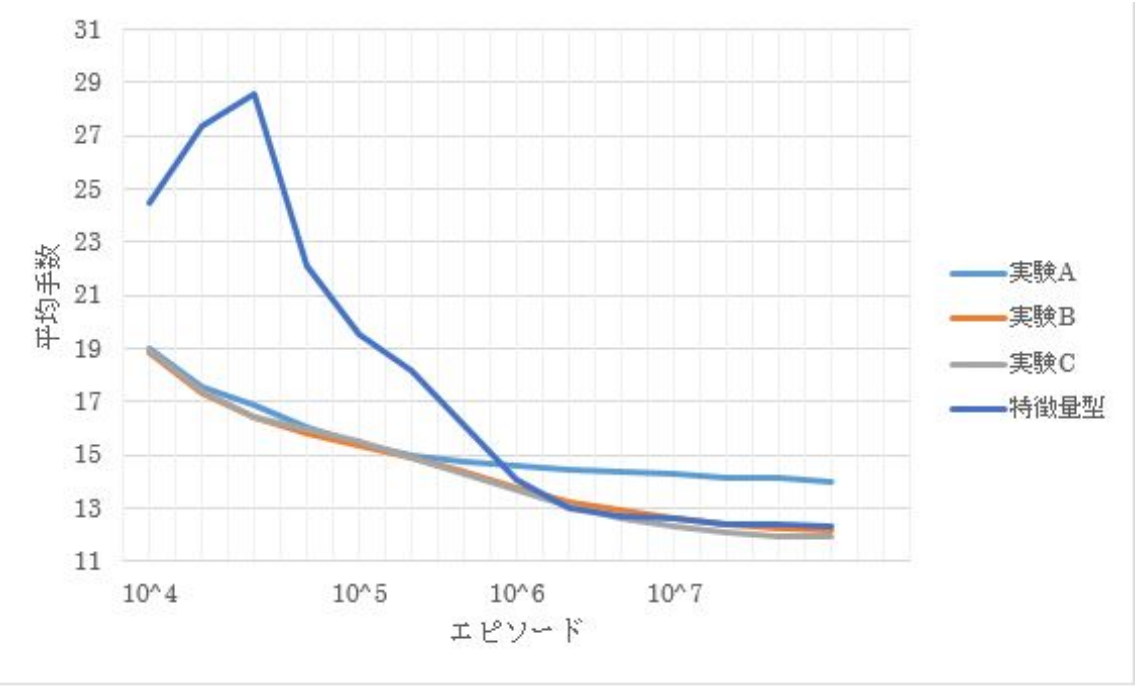

図 5.4: 2 色 5 枚麻雀の平均手数の比較

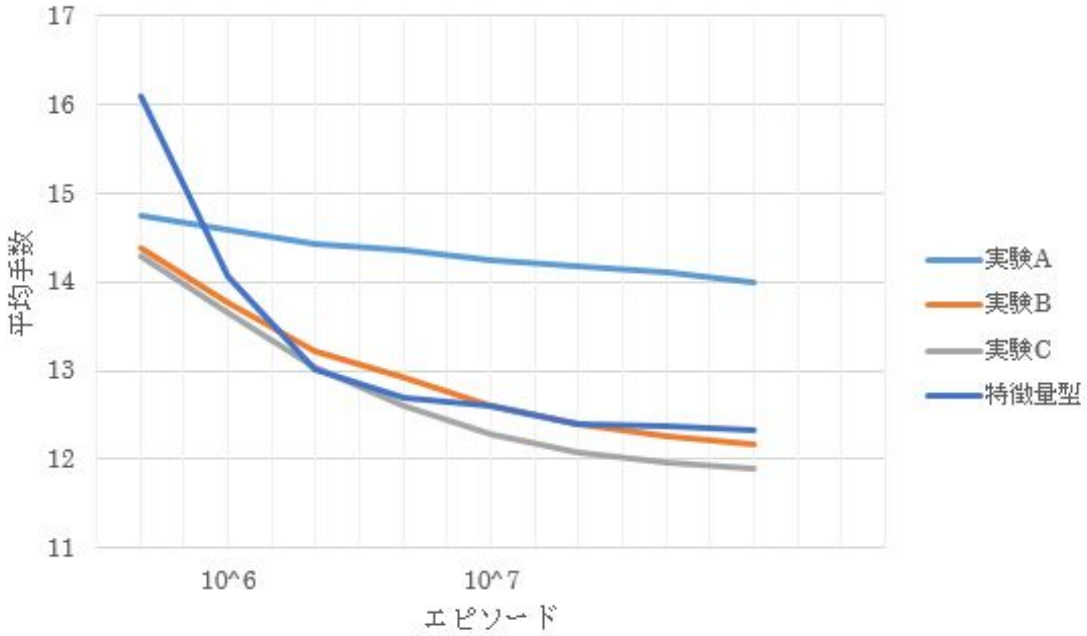

図 5.5: 後半拡大

表 5.6: テーブル型と特徴量型の比較

手法	和了までの平均手数	実験時間（秒）
実験 A	14.0	1253
実験 B	12.1	1189
実験 C	11.9	1116
特徴量型	12.3	12807

5.4　2 色 8 枚麻雀

3 枚麻雀から 1 色 5 枚麻雀へと手牌の数を増やした場合の実験と，1 色 5 枚麻雀から 2 色 5 枚麻雀へと牌の種類を増やした場合の実験は成功していると考えている．そこで，さらに複雑な問題として手牌の枚数を 8 枚に増やした 2 色 8 枚麻雀を設計した．和了の条件は雀頭 1 組と面子 2 組を揃えることである．また，麻雀の戦略にかかわる部分として，これまでの一定報酬の強化学習のほかに，和了時の獲得点数を報酬とする強化学習を行う．報酬を一定にした場合は和了までの平均手数が小さくなるような傾向が，獲得点数を報酬とした場合は和了時の平均手数が大きくなるような傾向が得られると予測する．そこで，固定報酬の場合を早アガリ，獲得点数報酬の場合を点数重視と呼称することにする．また，2 色 8 枚麻雀では常態の持たせ方が 27^7 と冗長となり，状態空間が大きくなってしまうため，状態をメモリに乗せることが困難である．よって，これ以降の単純化ゲームと 1 人麻雀ではテーブル型を用いていない．

実際の麻雀では和了したときの手牌にある役の翻の合計によって点数が決定される．2 色 8 枚麻雀は実際の麻雀よりも手牌の数が少なく，再現できる役は限られるため以下の役についてのみ実験を行う．また，実際の麻雀とは翻数や点数が異なるものもある．

1. メンゼンツモ　自分で引いた牌で和了する．この実験では和了時に必ず付与される（1 翻）
2. トイトイホー　2 つの面子がすべて刻子である（2 翻）
3. タンヤオ　各色の 1，9 牌を使用していない（1 翻）
4. イーペーコー　2 つの面子が同じ順子である（1 翻）
5. チンロートー　手牌のすべての牌が 1 か 9 の牌である（2 翻）
6. チンイツ　手牌のすべての牌が同じ色である（5 翻）

また，手牌の翻ごとの点数は表 5.7 の通りである．

表 5.7: 点数表

翻	点数
1	1000
2	2000
3	3900
4, 5	8000
6, 7	12000
8 以上	16000

そのほかの学習パラメータは以下の通りである．

- 割引率 $\gamma = 0.9$
- 総エピソード数 10^7
- $\alpha = 0.3 \times (10^4/10^4 + N)$　　$\alpha : 0.3 \to 3.0 \times 10^{-4}$
- $\varepsilon = 0.3$

2 色 5 枚麻雀に比べて手牌の数と特徴量の種類が増えたことで，1 エピソード当たりの実験時間も増大していると考えられる．そこで，総エピソード数を少なくして実験時間の短縮を図る．学習率の減少速度を変更しない場合は学習終了時までに学習率が十分小さくならないことが予想されたため，学習率の減少速度を速くした．

平均手数の結果は図 5.6 となった．どちらのプレイヤも学習が進むにつれて平均手数は少なくなっていった．実験終了時では早アガリの方が平均手数が約 3 手少ない．また，平均点数の結果は図 5.7 となった．点数重視のプレイヤは学習開始から平均点数が上昇し始め約 13500 点で収束している．一方で，早アガリのプレイヤでは，平均点数が横ばいになっている．

図 5.6，図 5.7 の結果から早アガリのプレイヤは平均手数が少なくなるような学習が進み，点数重視は早アガリよりも手数がかかる代わりに点数が高くなるような学習が進んだことがわかる．これらの結果から，与える報酬によって異なる戦略を得ることができたと考える．

実験で得られた平均手数，平均和了点数と実験にかかった時間を表 5.8 に示す．2 色 5 枚麻雀と比べると，1 エピソード当たりの実験時間は 2 倍程度に増えている．捨て牌を決定する際に 5 枚麻雀では手牌の中の 5 つの牌についてその牌を捨てたときの状態行動価値を求めるのに対して，8 枚麻雀では 8 枚の牌に対して計算すること，平均手数が 1.2 倍程度増えていることなどを踏まえると妥当であると考える．

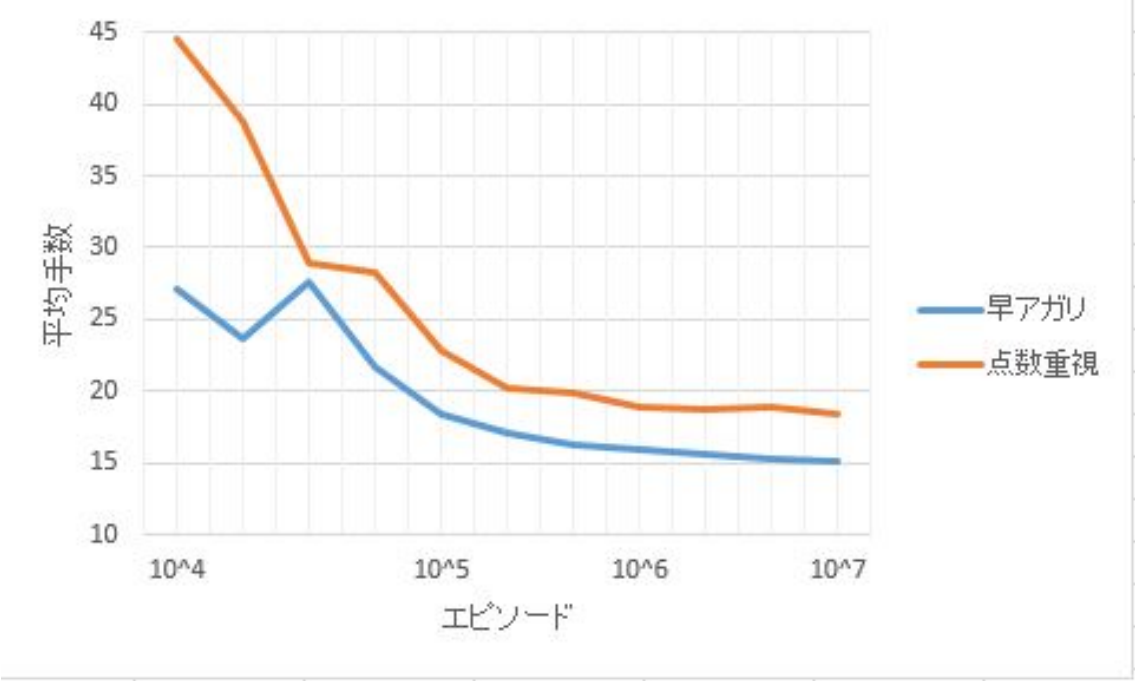

図 5.6: 平均手数の比較

表 5.8: 早アガリと点数重視の比較

手法	和了までの平均手数	平均和了点数	実験時間（秒）
早アガリ	15.2	6397	5070
点数重視	18.5	13448	6263

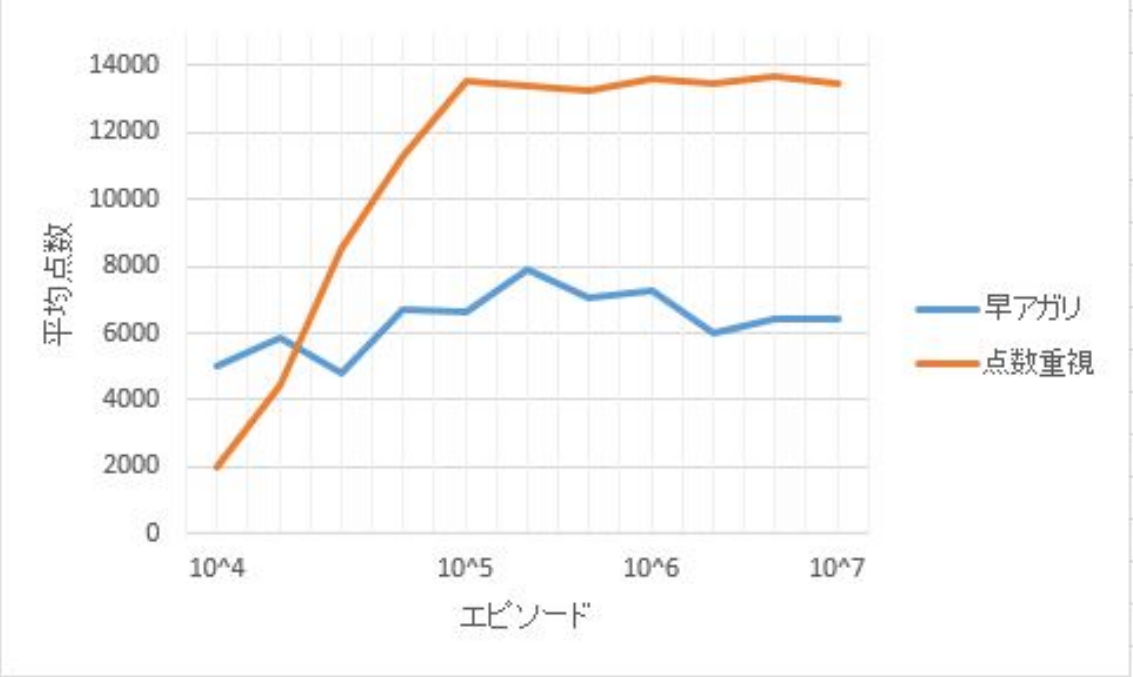
14000
12000
10000
8000
6000
4000
2000
0
平均点数
10^4
10^5
10^6
10^7
エピソード
早アガリ
点数重視

図 5.7: 平均点数の比較

5.5　3 色 8 枚麻雀

2 色 8 枚麻雀では報酬の与え方によって，異なる戦略を獲得することに成功していると考える．単純化ゲームの最後の実験として，獲得点数を報酬とした場合の強化学習に対して，割引率を変更した場合についての実験を行う．第 4.4.1 節で述べた通り，割引率を変えることで中間の戦略を得ることが予想される．割引率 0.6，0.7，0.9 の 3 種類の実験を行った．割引率の値が小さいほど，状態行動価値の 1 手辺りの割引が大きくなるため，平均手数が少ない戦略が得られるようになると予想される．また，一定報酬の早アガリの実験との比較も行う．3 色 8 枚麻雀は 3 種類の数牌を用いる．手牌の枚数や和了の条件，設定する役および点数表は 2 色 8 枚麻雀と同じものを用いる．

5.5.1　特徴量型実験

特徴量型の実験で用いる特徴量の種類は 112 種類である．そのほかの学習パラメータは以下の通りである．

- 総エピソード数 10^7
- 学習率 α: $0.3 \times \frac{10^4}{10^4+N}$ ，N は現在のエピソード数
  $(\alpha : 0.3 \to 3.0 \times 10^{-4})$
- $\varepsilon = 0.3$

平均手数の比較結果は図 5.8 の通りである．割引率 0.7 や 0.6 の場合は割引率 0.9 のプレイヤに比べて平均手数が少なく，早アガリのプレイヤより多くなっている．また，平均点数の比較結果は図 5.9 の通りである．割引率 0.7 や 0.6 の場合は割引率 0.9 のプレイヤに比べて平均点数が低く，早アガリのプレイヤより高くなっている．

表 5.9: 早アガリと点数重視の比較

手法	和了までの平均手数	平均和了点数	実験時間（秒）
早アガリ	24.3	4312	8602
割引率 0.9	29.8	11068	9185
割引率 0.7	26.2	6733	9051
割引率 0.6	26.1	6586	8995

この結果から，1 手ごとの割引が大きくなるように割引率を変更することで，点数重視に比べて平均手数が少なく，獲得点数が低い中間の戦略が得られたと考える．

実験時間は 2 色 8 枚麻雀に比べて 1.5 倍程度長くなった．平均手数が約 1.7 倍，特徴量の種類も 36 種類増えていることが原因だと考える．

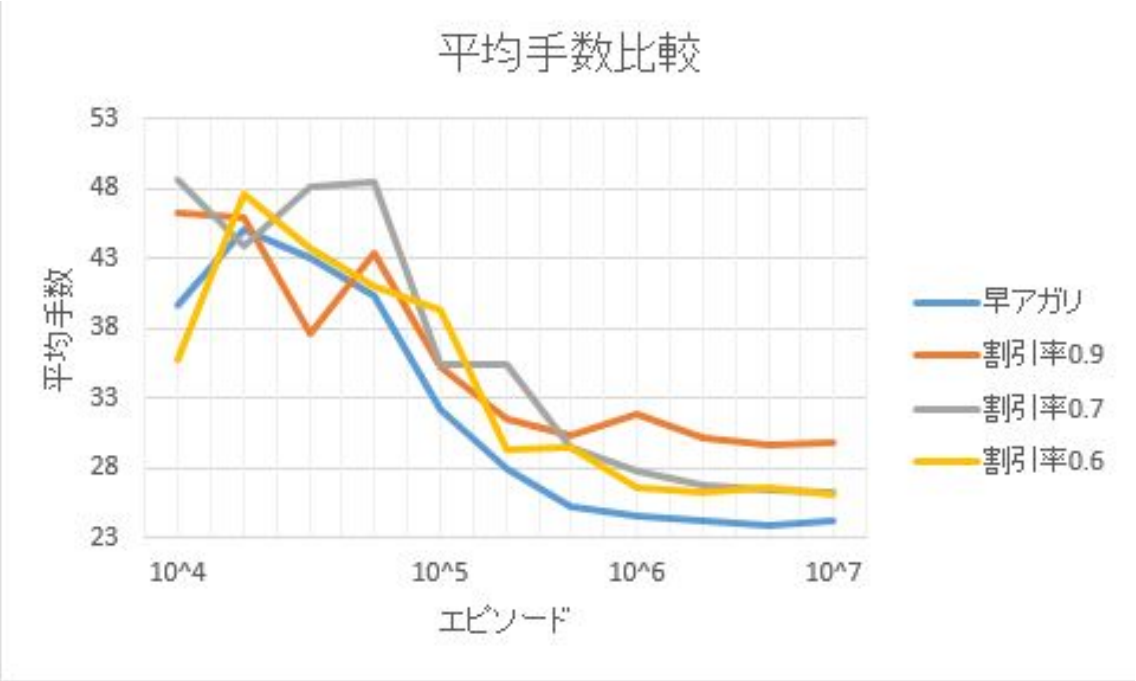

図 5.8: 平均手数の比較

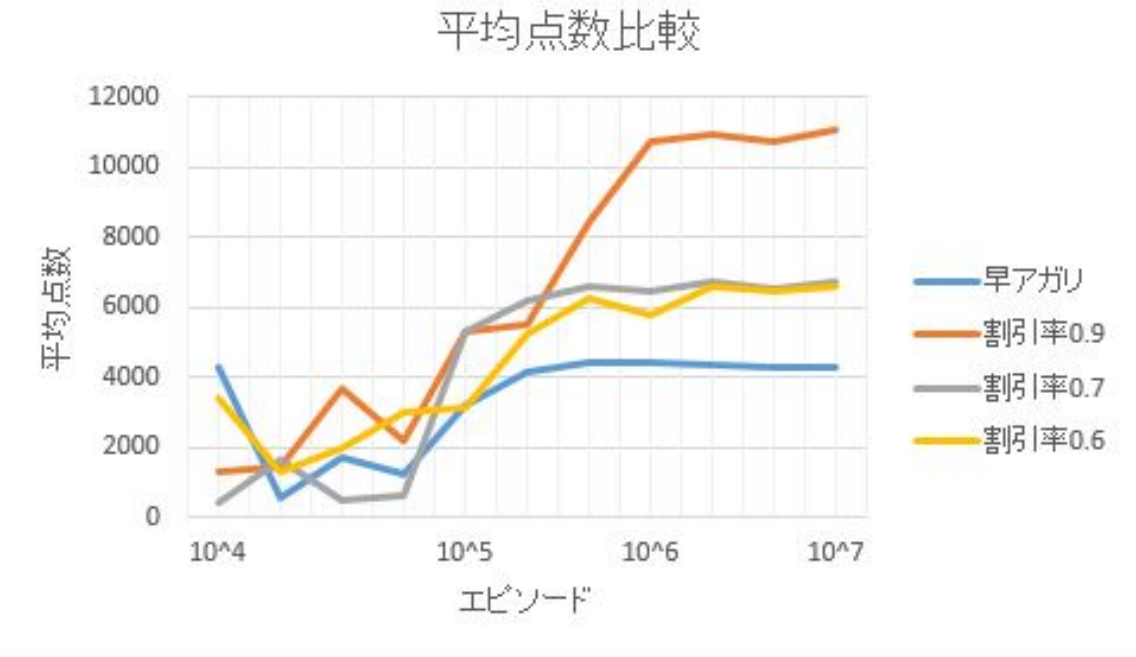

図 5.9: 平均点数の比較

5.6 単純化ゲーム問題まとめ

本研究で実装・実験した単純化ゲームに対する研究結果をまとめる．

- 5.1 節では 3 枚麻雀について，Q（状態，行動）のタイプの強化学習を行った．テーブル型と特徴量型について，すべての状態行動価値が理論値に収束していることを確認した．
- 5.2 節では 1 色 5 枚麻雀について，Q（次状態）のタイプの強化学習を行った．テーブル型では特定の状態行動価値が理論値に収束していることを確認した．特徴量型では，状態行動価値の理論値こそ得られていないものの，手数は同じところに収束している．
- 5.3 節では，2 色 5 枚麻雀について，学習率のスケジューリングが大事であることを確認した．テーブル型では状態数が多くなると各状態の経験数が足りなくなりがちであることを確認した．
- 5.4 節では 2 色 8 枚麻雀について，テーブル型については学習が困難であるという判断から実験をあきらめ，特徴量型を用いた．和了時の手牌に対する獲得点数を定義し，「早アガリ」「点数重視」の 2 つの強化学習を行った．
- 5.5 節では 3 色 8 枚麻雀について，さらにこの 2 つの中間の戦略を得るべく割引率を調整した．

6 章では，1 人麻雀に取り組む．上記からさらに「手数制限」「手数グルーピング」などの概念が追加され，手法として新たにニューラルネットワークを用いたものが適用される．

第6章　1人麻雀

単純化ゲームで得られた知見をもとに 1 人麻雀に対して特徴量型・ニューラルネット型の強化学習を適用する．一定報酬による早アガリのプレイヤ，点数報酬による点数重視のプレイヤ，そして割引率を調整して，これらの中間の戦略を持つプレイヤの獲得を目指す．

6.1　特徴量型実験

6.1.1　実験設定

特徴量は 192 種類とした．また，海津らの研究 [6] を参考にして，本研究の 1 人麻雀に用いる役を決定した．本研究の 1 人麻雀に用いる役は表 6.1 の通りである．

役牌は，白，發，中の 3 つの牌と局開始時の自分の風と同じ自風牌，局開始時の場の風と同じ場風牌がある．こちらも，海津らの研究 [6] を参考にして自風，場風を共に東とした．また，局開始時に 34 種類の中からランダムに 1 種類の牌を選びドラとした．ドラは和了時に手牌にある枚数と同じ数の翻を得ることができる牌である．

点数表は表 6.2 の通りである．

ゲームの最大手数は，学習時では，制限を設けるといつまでも和了できず学習が進まない可能性が高いことを考慮し，制限を設けていない．一方で，テスト時は良い戦略が得られていない学習初期において，和了できないような重みを獲得した状態で捨て牌の選択を行ってしまい，いつまでも和了できないという可能性があるため 50 とした．この場合，学習時は何回でも牌が引けることを前提とした戦略を得てしまい，良い戦略が得られないことが懸念される．そこで，学習時のエピソードが 1 万エピソードに達するまでは手数に制限を設けず，それ以降の学習ではテスト時と同じように制限を設けるという設定とした．

表 6.1: 1 人麻雀の役

役名	翻数
メンゼンツモ	1
平和	1
タンヤオ	1
役牌	1
小三元	2
三暗刻	2
混老頭	2
三色同刻	2
対々和	2
一気通貫	2
三色同順	2
全帯	2
二盃口	3
混一色	3
九連宝燈	13
大四喜	13
字一色	13
緑一色	13
小四喜	13
大三元	13
清老頭	13

表 6.2: 1 人麻雀の点数表

翻数	点数
1	1000
2	2000
3	3900
4	7700
5	8000
6,7	12000
8,9,10	18000
11,12	24000
1	32000

手数重視（一定報酬），点数重視（獲得点数報酬）について実験し，獲得点数報酬の割引率を変更して中間戦略が得られるかを実験した．学習パラメータは以下の通りである．

- 総エピソード数 10^7
- 学習率 α: $0.3 \times \frac{10^4}{10^4+N}$ ，N は現在のエピソード数
  $(\alpha : 0.3 \to 3.0 \times 10^{-4})$
- $\varepsilon = 0.1$

手牌と牌の種類が多い 1 人麻雀では，ランダムな行動を多くとってしまうと和了することができなくなる可能性が考えられる．そこで，単純化ゲームの実験のときよりも ε の値を小さくしてランダムに行動する確率を小さくする．

また，各プレイヤの割引率は次の通りである．

- 手数重視：$\gamma = 0.9$
- 点数重視：$\gamma = 0.9$
- 中間戦略：$\gamma = 0.7$

6.1.2 実験結果 1

実験で得られた学習曲線は図 6.1，図 6.2 の通りである．図 6.1 から，手数重視プレイヤと中間戦略のプレイヤで学習が進むにつれて平均手数が減少していることがわかる．点数重視のプレイヤは平均手数の減少はそこまで大きくはないが，平均点数は他のプレイヤよりも高くなっている．また，割引率 0.7 のプレイヤは手数重視よりも点数が高く，点数重視よりも手数の少ない中間の戦略が得られている．これらのことから，報酬や学習パラメータの調整によって異なる戦略を得られたと考える．

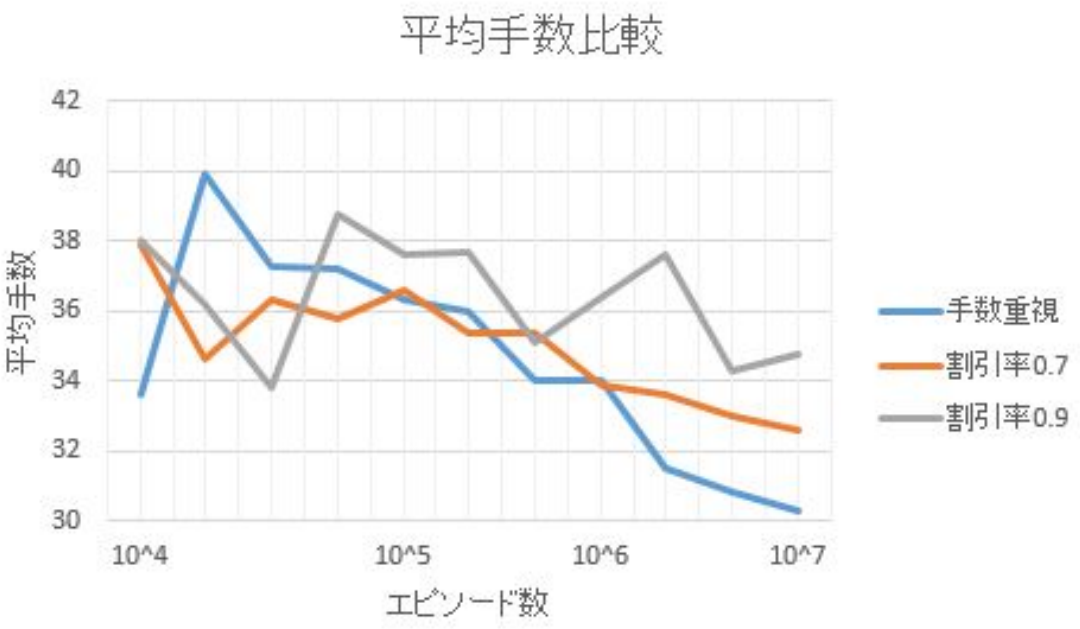

図 6.1: 平均手数の比較

表 6.3: 各プレイヤの比較

手法	和了までの平均手数	平均和了点数	実験時間（秒）
手数重視	30.3	3712	51105
割引率 0.7	32.6	4123	53341
割引率 0.9	34.7	7991	57196


平均点数比較結果
14000
12000
10000
8000
6000
4000
2000
0
平均点数
10^4
10^5
10^6
10^7
エピソード数
手数重視
割引率0.7
割引率0.9

図 6.2: 平均点数の比較

6.1.3 具体的な戦略例

手数重視のプレイヤと点数重視のプレイヤについて，実際に同じ手牌を与えてどのように捨て牌を選択するのかを調べた．手牌として図 6.3 を与える．手数重視のプレイヤでは，9 ピンを捨てる手が選ばれた（図 6.4 参照）．これは，4 ピン，6 ピン，7 ピンで和了することのできる聴牌であり，和了時の点数は 4 ピンか 7 ピンを引いたときが 1000 点，6 ピンを引いたときが 2000 点である．一方で，点数重視のプレイヤでは，4 ソウを捨てる手が選ばれた（図 6.5 参照）．これは向聴数が 2 であるが，9 ピンを残すことで清一色などの高い役を狙うことができる手である．

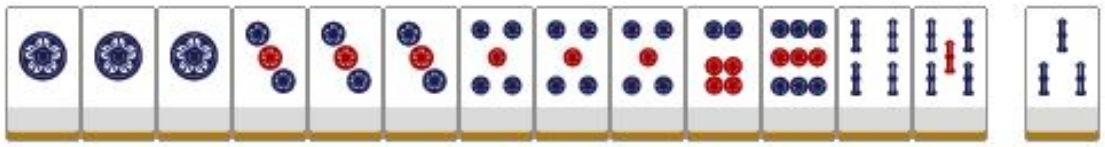

図 6.3: 得られた戦略の比較

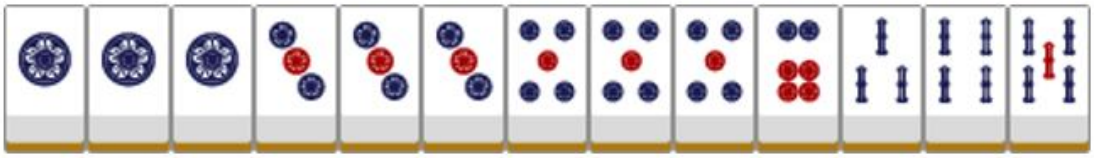

図 6.4: 手数重視の待ち方

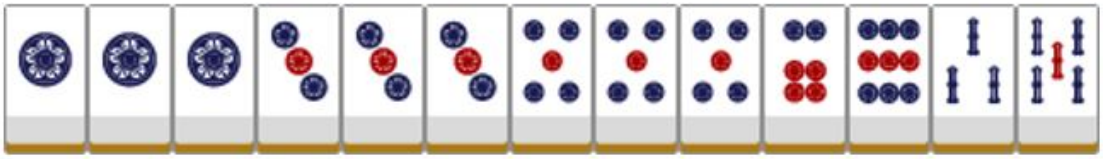

図 6.5: 点数重視の待ち方

6.1.4 既存研究との比較

本研究で得られたプレイヤと，海津らの研究 [6] で得られた 1 人麻雀プレイヤのうち，早アガリを目指すプレイヤと点数の高いアガリを目指すプレイヤと比較を行う．海津らの研究では山を設定していること，最大手数を 18 手に制限していることが本研究と異なる．そこで，比較では本研究で得られたプレイヤが 18 手以内に和了した場合の和了確率，平均手数，平均点数を用いる．また，海津らの研究では得られる点数を親の点数（親は子に比べて 1.5 倍程度の点数を得る）を用いていたので，実験で得られた翻数をもとに比較しやすいように本実験で得られた点数を調整している．

表 6.4: 既存研究との比較結果

手法	アガリ確率	和了時の平均点数	和了時の平均手数	テスト回数
本実験手数重視	11.5 %	3440.0	13.9	10^6 回
本実験点数重視	6.5 %	7742.4	14.1	10^6 回
既存研究手数重視	17.5 %	3924.0	14.0	10^4 回
既存研究点数重視	11.3 %	8022.4	14.9	10^4 回

表 6.4 より，平均手数は既存研究と同等以上の結果が出ているが，一方でアガリ確率や平均点数では既存研究に劣っている．このことについて，既存研究では学習時から 18 手で和了することを前提とした学習をしていたのに対し，本実験では 50 手を最大手数としてしまったことが，十分な結果を得られなかった原因の 1 つであると考察する．

また，麻雀では牌を多く引くことができる局の序盤と，引ける回数が少なくなる終盤ではとるべき戦略が異なる．そこで，手数ごとに特徴量の重みを設定し，現在の手数に応じた重みを用いて捨て牌の選択を行う手法を考える．18 ある手数 1 つずつに対して重みを設定してしまうと，1 エピソードで 1 回ずつしか重みの更新が行われず，重みの収束に膨大な時間を要することが考えられるため，本研究では 18 の手数を序盤，中盤，終盤の 3 つに分割し，それぞれに対して重みを設定した．この手法を手数グルーピング手法と呼称する．

6.2 手数グルーピング手法実験

麻雀において多くの牌を引くことができる局の序盤と，引くことのできる回数が少ない局の終盤では異なる戦略をとるべきである．そこで，1 局の序盤，中盤，終盤についてそれぞれ異なる重みを用いた特徴量型の手法を提案する．1 局の最大手数を 18 として 1 から 6 手までを序盤，7 から 12 手までを中盤，13 から 18 手を終盤としてそれぞれに異なる重みを用いた実験を行う．例えば，現在の手数が 9 手目であれば中盤の重みを用いて状態行動価値を計算して手の選択を行い，中盤の重みを更新する．この手法を用いることで，点数重視のプレイヤでも終盤は他方点数が安くなっても和了を目指すなど和了確率の向上が期待できる．また，最大手数を設けることで実験時間の短縮も行えると考える．

学習パラメータは以下の通りである．

- 割引率 $\gamma = 0.9$
- 報酬 一定報酬（手数重視）　和了時の獲得点数（点数重視）
- 総エピソード数 4.0×10^7
- 学習率 α: $0.3 \times \frac{10^4}{10^4+N}$ ，N は現在のエピソード数
  ($\alpha : 0.3 \to 3.0 \times 10^{-4}$)
- $\varepsilon = 0.3$

和了確率の学習曲線を図 6.6 に示す．また，手数グルーピング手法の実験で最終的に得られた結果と手数グルーピングを行わなかった特徴量型の手法との比較を表**??**にまとめた．手数重視では 18 手以内の和了確率はほとんど増加しなかったが，点数重視では約 1.5 倍と大きく増加している．一方で，平均点数は手数グルーピングありの方が低くなっている．これは，終盤に多少点数が低くなったとしても和了しようとする戦略が学習された結果，全体の平均点数が下がったためであると考える．また，実験時間は約 $\frac{1}{3}$ まで短くなっている．最大手数が設けられたことで，実験全体の手数が減少したことが原因と考える．

表 6.5: 手数グルーピングの有無の比較結果

手法	アガリ確率	和了時の平均点数	和了時の平均手数	テスト回数	実験時間（秒）
手数重視 （手数グルーピングあり）	11.7 %	3026.2	13.9	10^6 回	15800
点数重視 （手数グルーピングあり）	9.6 %	5075.8	13.8	10^6 回	18762
手数重視 （手数グルーピングなし）	11.5 %	3440.0	13.9	10^6 回	51105
点数重視 （手数グルーピングなし）	6.5 %	7742.4	14.1	10^6 回	57196

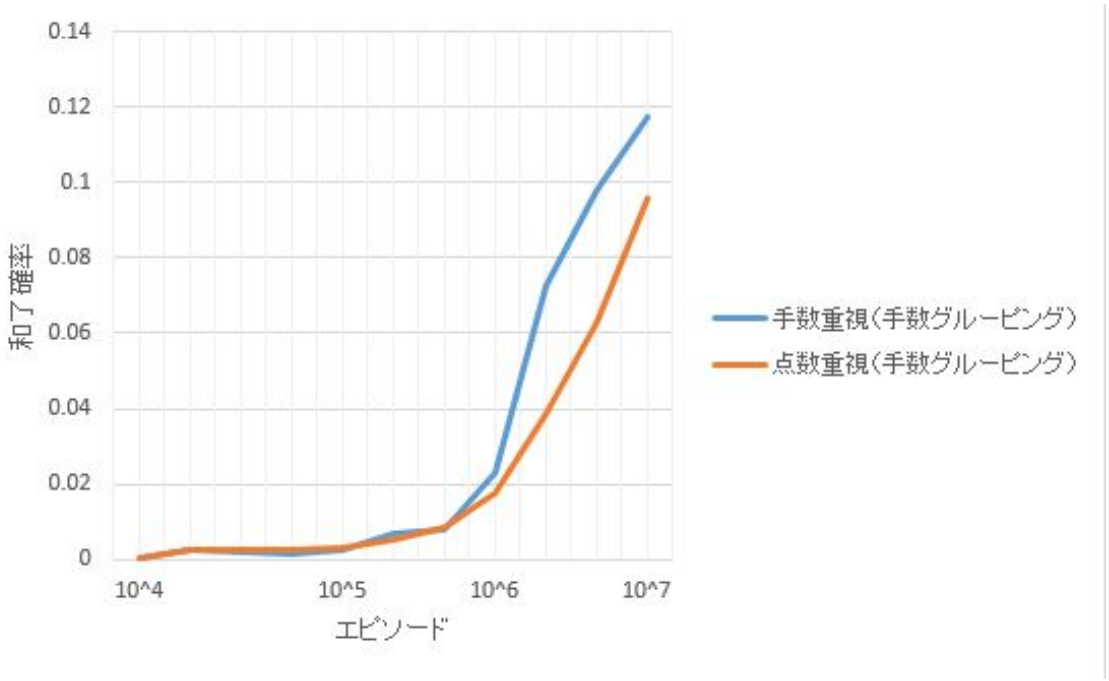

図 6.6: 手数グルーピング和了確率

6.3 ニューラルネット型

6.3.1 実験の目的

手数グルーピング手法を用いた場合の和了確率は既存研究に劣っていた．このことから，1 人麻雀の正確な状態行動価値や行動の選択は得られていないと考える．この理由の 1 つとして，特徴量型のモデルでは状態行動価値に対する表現力が不足していることを挙げる．より表現力の高いモデルを用いることで性能の向上ができると期待する．特徴量型では特徴量の重みに対して単純な線形和を用いることで状態行動価値を表現していた．そこで，より複雑なモデルとして，ニューラルネットワークで状態行動価値を近似する“ニューラルネット型”を 1 人麻雀に適用する．

6.3.2 実験設定

特徴量型では 1 ゲームごとに状態行動価値の更新が行われたが，ニューラルネット型で同じように 1 ゲームごとに更新を行うと実験時間の増大や汎化性能の悪化が考えられる．そこで，ある程度まとまった回数のゲームを行い，そのデータをもとにニューラルネットワークの更新を行う．ゲームを行い価値を更新するまでを，本研究ではサイクルと呼称する．また，ε-greedy 法によって，ランダムな手をとり和了できなかった場合，実際の状態行動価値よりも低い値をデータとして保存してしまう欠点がある．そこで，ゲームの回数が進むごとに ε の値を小さくすることで，学習後半に低すぎる状態行動価値が学習されにくい設定とする．

この手法のもう 1 つの問題点として，学習を行う前の最初の 1 ゲームの行動選択をどうするのかという問題がある．完全にランダムに行動してしまうと，1 人麻雀ではほとんど和了できないため指標となる価値を学習できない．一方で，今までの特徴量型の実験で学習した状態行動価値によって得られた学習データを与えると，学習データがある程度高い性能を持っているために，正確ではない 2 サイクル目以降がよりよい価値の更新を行うことができない可能性がある．そこで，1 サイクル目では今までの特徴量型の学習によってデータを収集するが，その学習の際にその後のニューラルネット型で和了したときに得られる報酬よりも少ない報酬を与えることを提案する．これにより，1 サイクル目のデータがその後の学習に与える影響を抑えることができると考える．

ニューラルネットワークの学習はバッチと呼ばれるデータの一部を用いた学習を繰り返して行う機械学習である．この繰り返し回数をエポック数と呼ぶが，エポック数が多くなりすぎると，過学習を起こし汎化性能が下がることが知られている．そこで，エポック数ごとに汎化性能が低下したかどうかを調べ，減少した場合はそこで学習を打ち切る Early-Stopping という手法を導入した．

ニューラルネットワークを以下のように設定した．ネットワークの実装には python3.5 の Keras ライブラリを用いており，最適化の設定として「adam」を用いている．

- 入力層のニューロン数　192
- 中間層のニューロン数　5
- 出力層のニューロン数　1
- 隠れ層の発火関数　ReLU 関数

学習パラメータは以下の通りである．また最初に与えるデータとして，6.1 節の特徴量型，手数重視の設定で報酬のみ $\frac{1}{5}$ としたプレイヤを学習させ，最終的に得られたプレイヤで 10^4 ゲームプレイしたデータを与える．

- 学習率 α: $0.05 \times \frac{10^4}{(10^4)*C}$ ，C は現在のサイクル数
  ($\alpha : 0.05 \rightarrow 0.005$)
- 割引率 $\gamma = 0.9$
- 2 サイクル目以降　和了したときに 100 を与え，それ以外では報酬は与えない
- 1 ゲームの最大行動回数　50 手
- 1 サイクル当たりのゲーム数　10^4 ゲーム
- 総サイクル数 10 サイクル（学習で行う総ゲーム数 10^5ゲーム）
- $\varepsilon = 0.1 \times 0.8^C$ ，C は現在のサイクル数
  ($\varepsilon : 0.1 \rightarrow 0.01$)

6.3.3 実験結果

実験結果は図 6.7 の通りである．9 サイクル目まで和了確率は増加し続けている．また，最大手数を 50 手とした場合の特徴量型との比較を表 6.6 に示す．特徴量型（途中）は 10^5 ゲームの学習が終了した時点の性能を示し，特徴量型（最終）は学習がすべて終了した時点の性能を示している．本研究の設定のニューラルネット型では特徴量型に比べてアガリ確率は向上しなかった．一方で，同じデータ数を与えた場合の比較ではニューラルネット型の方がよい結果を示している．ニューラルネット型はデータ数を少なくした場合でもよい結果が得られる点が利点であるといえる．本研究の 1 人麻雀では学習に用いるデータ数をそろえることにはあまり意味がないが，現実の問題ではデータ数の方が計算時間よりも重要となる問題は存在しうるため，少ないデータ数で学習できたことには一定の意義があると考える．

1 ゲーム当たりの実験時間は特徴量型に比べて約 30 倍となっている．状態行動価値の計算にニューラルネットワークを用いており，線形和の特徴量型に比べて計算時間が長くなったことが原因だと考える．

表 6.6: ニューラルネット型と特徴量型の比較結果

手法	アガリ確率	学習に用いた総ゲーム回数	実験時間（秒）
ニューラルネット型	72.1 %	10^5	14455
特徴量型（途中）	56.0 %	10^5	511
特徴量型（最終）	82.2 %	10^7	51105

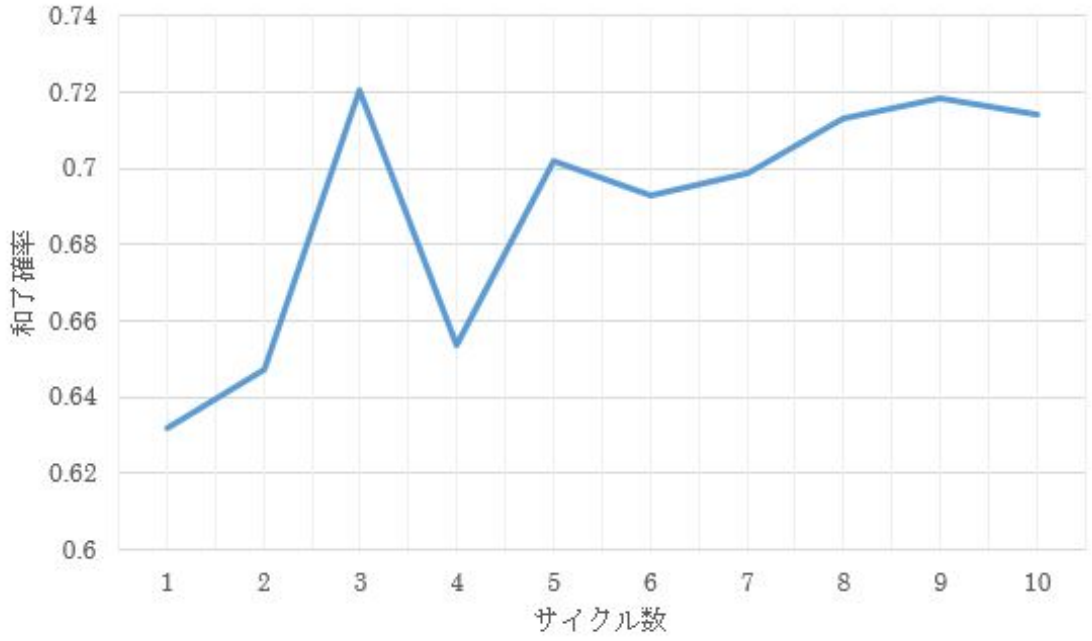

図 6.7: ニューラルネット型和了確率

第7章　まとめ

本研究では，麻雀における多様な戦略を容易に得るために，相手プレイヤがいないものとして捨て牌の選択を行い和了を目指す 1 人麻雀と，それを単純化したゲームに対して“テーブル型”“特徴量型”“ニューラルネット型”の 3 種類の強化学習を適用した．状態空間の小さい簡単な問題に対しては，テーブル型を用いることで特定の状態行動価値をほぼ正確に学習することができたと考えている．特徴量型では，1 人麻雀の報酬を切り替えることで,「早アガリ」と「点数重視」の 2 種類のプレイヤの作成を行い，異なる戦略を学習することができた．また，点数重視プレイヤの割引率を調整することにより，それらの中間の戦略をとるプレイヤを獲得した．1 人麻雀では，1 局の序盤，中盤，終盤ごとに異なる特徴量の重みを用いることで，和了確率の性能向上を行うことができた．ニューラルネット型では最終的な性能は特徴量型に劣ったものの，学習に同じゲーム数を与える場合はニューラルネット型の方が性能がよいため，サイクル数や 1 サイクル当たりのデータ数を増やすことで性能の向上が見込めると考えている．

今後の展望として，特徴量の種類を吟味することによる性能向上や，順位を上げることを目的とするために，一定以上の獲得点数を得た場合の報酬を増やした強化学習などを行うことで，より実際の麻雀の多様な状況に適した戦略を獲得することなどがある．また，得られたプレイヤを実際に人間プレイヤや既存のコンピュータプレイヤと対戦させることでの評価も考えられる．

謝辞

本研究を進めるにあたり，研究の機会を与えていただき，終始熱心にご指導していただいた情報科学研究科池田心准教授に深謝いたします．また，技術指導をしてくださった佐藤直之さんをはじめ，研究生活を支えてくださった池田研究室のすべての方々に感謝いたします．

参考文献

[1] Campbell Murray A. Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. Artificial intelligence, 134.1 pp.57-83, 2002.

[2] 保木邦仁. 局面評価の学習を目指した探索結果の最適制御. 第 11 回ゲームプログラミングワークショップ, pp.78-83, 2006.

[3] David Silver, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 529.7587 pp.484-489, 2016.

[4] 池田心，楽しませる囲碁･将棋プログラミング，オペレーションズ･リサーチ学会，Vol.58，no.3，pp167-173.（2013）

[5] 水上直紀，鶴岡慶雅，期待最終順位に基づくコンピュータ麻雀プレイヤの構築，The 20th Game Programming Workshop 2015，pp.179-186,（2015）

[6] 海津純平，成澤和志，篠原歩，一人麻雀における打ち方を考慮した評価指標に関する研究，The 20th Game Programming Workshop 2015，pp.172-178,（2015）

[7] Sutton, R. S. and Barto, A.: Reinforcement Learning: An Introduction, A Bradford Book, the MIT Press (1998)

[8] 水上直紀，中張遼太郎，浦晃，三輪誠，鶴岡慶雅，近山隆，降りるべき局面の認識による 1 人麻雀プレイヤの 4 人麻雀への適用

[9] 角田真吾．天鳳，http://tenhou.net/（アクセス日時：2018.01.29）

[10] 三木理斗，多人数不完全情報ゲームにおける最適行動決定に関する研究，修士論文，東京大学,（2010）

[11] 田中悠，池田心，麻雀初心者のための状況に応じた着手モデル選択，第 31 回ゲーム情報学研究会，pp.1-8,（2014）

[12] 築地毅，柴原一友，ディープラーニング麻雀－オートエンコーダとドロップアウトの有効性－，The 19th Game Programming Workshop 2015．pp.136-142,（2015）

[13] 萩原涼太，山田渉央，佐藤直之，池田心，麻雀における相手の和了点数予測法の性能評価，第 35 回ゲーム情報学研究会，pp.1-8,（2016）

[14] 水上直紀，鶴岡慶雅，牌譜を用いた対戦相手のモデル化とモンテカルロ法によるコンピュータ麻雀プレイヤの構築，The 19th Game Programming Workshop 2014，pp.48-55,（2014）

[15] 佐藤直之，池田心，花札のこいこいにおける方策勾配法と Neural Fitted Q Iteration の適用，The 22th Game Programming Workshop 2017．pp.64-71,（2017）

[16] Martin Riedmiller, Neural fitted Q iteration-first experiences with a data efficient neural reinforcement learning method. Lecture Notes in Computer Science: European Conference on Machine Learning, pp.317328(2005)