令和 5 年度

修士学位論文

合議アルゴリズムによる AlphaDDA の動的強さ調整の改良

Improving Dynamic Strength Adjustment of AlphaDDA with Ensemble Algorithms

1265099　　久保田　留奈

指導教員　　松崎公紀

2024 年 2 月 28 日

高知工科大学大学院 工学研究科 基盤工学専攻

情報学コース

要 旨

合議アルゴリズムによる AlphaDDA の動的強さ調整の改良

久保田　留奈

ゲーム情報学の分野では，人間の練習相手として AI を活用することが新たな目標となっている．人間相手にプレイする場合，AI プレイヤーが極端に強い手・弱い手を打つと人間は不快を感じるため，対戦相手の強さに合わせて AI プレイヤーの強さを適切に調整すること（動的強さ調整）が重要である．

Fujita が 2022 年に提案した AlphaDDA（AlphaZero により得られる強いプレイヤーに対し，動的な強さ調整を導入した AI）では，対局においておよそ 50%の勝率を達成することができた．しかし，対局中の局面の評価値から再評価してみると評価値が極端に高いもしくは低い瞬間があった．これは，AlphaDDA の強さがゲーム全体を通して適切に制御できていなかったことを示す．

そこで本研究では，複数の思考プログラムの候補手から一つの手を選択させる合議と呼ばれる手法を用い，より互角に近い手を選択させることで動的強さ調整の改良を目指した．多数決合議と楽観合議による合議プレイヤーの対局結果では，単一の DDA プレイヤーと比較すると極端な評価値の振れは減少したが勝率の面などでは課題が残った．

キーワード　動的強さ調整，AlphaZero，合議

Abstract

Improving Dynamic Strength Adjustment of AlphaDDA with Ensemble Algorithms

KUBOTA, Runa

Using AI as a practice partner for humans has been a new goal in game informatics. When playing against humans, called appropriately adjusting the AI player's strength to match that of the opponent (Dynamic Strength Adjustment, or DSA) is crucial. If an AI player in consistently makes extremely strong or weak moves, it can lead to discomfort for human players.

AlphaDDA (AI with dynamic strength adjustment for strong players obtained by AlphaZero) proposed by Fujita in 2022 enabled players to achieve an approximate 50% win rate. However, re-evaluating the game based on position evaluation values during games revealed that there were moments of significantly high/low evaluation values, indicating that the AlphaDDA's strength was not adequately controlled throughout.

This study aimed to improve dynamic strength adjustment by using a method called "Ensemble Algorithms," which involves selecting a single move from multiple candidate moves generated by different programs.

In matches with ensemble algorithms player using majority voting and optimistic, extreme fluctuations in evaluation values were reduced compared to a single DSA player. However, there were remained challenges for controlling win rate.

key words Dynamic Strength Adjustment, AlphaZero, Ensemble Algorithms

目次

図目次

図目次

表目次

第 1 章

はじめに

近年，人工知能の発展は目覚ましく，ゲーム分野では AlphaZero のように人間のトッププロを超える強さの AI が開発されている．またそのような人間を超える強さを持つ AI を，人間の練習相手のために用いることを目標とした研究 [1] も行われている．人間相手にプレイする場合，AI プレイヤーが極端に強い手・弱い手を打つと人間は不快を感じるため，対戦相手の強さに合わせて AI プレイヤーの強さを適切に調整する動的強さ調整が重要である．

AlphaZero により得られた強いプレイヤーに動的強さ調整を導入した AI に，Fujita が提案した AlphaDDA[1] がある．AlphaDDA プレイヤーは対局においておよそ 50 %の勝率を達成できていた．しかし，各局面における局面評価値の変動による再評価を行った [2] 結果，実際は局面評価値は大きく振れる場面があり，ゲーム全体に渡って適切に強さを制御できているわけではないことがわかった．

そこで本研究では，複数の思考プログラムの候補手から一つの手を選択する合議と呼ばれる手法を用い，より互角に近い手を選択させることで動的な強さの調整が可能か実験を行なった．結果として，合議プレイヤーでは単一の DDA プレイヤーに比べると局面評価値の変動の幅を概ね小さくすることができた．しかし，勝率は単一のプレイヤーより下がるという結果であった．

本論文は次のとおりの構成である．第 2 章では，AlphaDDA 及び使用した動的強さ調整手法について説明する．第 3 章では，AlphaDDA の再評価として各 DDA プレイヤーの対局中の局面評価値の変動について考察する．第 4 章では，DDA プレイヤーの合議について，第 5 章で合議実験の結果を記述する．第 6 章で，強さ調整に関して関連のある研究を説明し，第 7 章で本論文をまとめる．

第 2 章

AlphaDDA

動的強さ調整手法に，Fujita による AlphaDDA[1] がある．AlphaDDA は，AlphaZero により得られる強いプレイヤーに対し，3 種類の動的な強さの調整を導入した AI プレイヤーである．

AlphaDDA のベースである AlphaZero について簡単に説明する．AlphaZero は，ディープニューラルネットワーク（DNN）とモンテカルロ木探索（MCTS）から構成されている．DNN では，局面状態 s に対して評価値 $v(s)$ と次の手の選択確率を予測する．入力は，現在のプレイヤーの局面の状態とプレイヤーの先手・後手の情報である．MCTS は，ゲームの可能な手の木を効率的に探索し，最適な手を見つけるために使用する．AlphaZero における MCTS については付録に記載した．

AlphaDDA は，学習させた AlphaZero に動的強さ調整を加えたプレイヤーであり，AlphaZero と同様に DNN と MCTS から構成される．DNN において AlphaDDA の勝率が高いと予測した場合弱い手を選択し，AlphaDDA の勝率が低い場合強い手を選択する．

AlphaDDA は，DNN から局面の評価値を取得する．DNN は，局面の状態から評価値と手の選択確率を推定している．n ターン目の評価値 V_n は，-1〜1 の範囲であり，直前 N_h 個の評価値の平均で $\bar{V}_n$ は定義される．

$$\bar{V}_n = \frac{1}{N_h} \sum_{i=0}^{N_h-1} (V_{n-i})$$

AlphaDDA は，この $\bar{V}_n$ の値に従って強さを調整することで動的強さ調整を行う．

Fujita によって提案された AlphaDDA はそれぞれ，MCTS のシミュレーション回数

（DDA-M）・ドロップアウト確率（DDA-D）・UCT スコア（DDA-U）の調整を行ったプレイヤーである．さらに本研究では，softmax 方策を用いた強さ調整手法（DDA-S）[4] を加えた 4 種類の DDA プレイヤーを使用する．

以下で各 DDA プレイヤーについて説明する．

2.1　DDA-M プレイヤー

DDA-M は，MCTS におけるシミュレーション回数を変更することで DDA を実現する手法を用いている．MCTS ではシミュレーション回数が多いほど強い手を選択する精度が向上するため，シミュレーション回数が多くなればより強い手，少なくなればより弱い手を選択させることができる．

評価値 $\bar{V}_n$ におけるシミュレーション回数 N_{sim} を以下に示す．なお，N_{sim} は 1 以下は 1，400 以上は 400 の値としている．$player$ は，先手プレイヤーが 1 後手プレイヤーが -1 である．

$$N_{sim}(\bar{v}_n) = \lceil 10^{-2.8(\bar{v}_n * player - 1.4)} \rceil$$

2.2　DDA-D プレイヤー

DDA-D は，DNN 内の一部を無視する，ドロップアウトの確率を変更することで DDA を実現する手法を用いている．ドロップアウトの確率が小さいほど，学習させた AlphaZero と同等のプレイヤーとなるため強い手を選択し，ドロップアウトの確率が大きいほど弱い手を選択するプレイヤーとなる．

評価値 $\bar{V}_n$ におけるドロップアウト確率 P_{drop} を以下に示す．なお，P_{drop} は 0 以下は 0，0.95 以上は 0.95 としている．

$$P_{drop}(\bar{v}) = 20(\bar{V}_n - 0.9)$$

2.3 DDA-U プレイヤー

DDA-U は，評価値に応じて C を変更して MCTS における UCT の値を変化させるプレイヤーである．C では未知の手の探索を優先するか，既知の良い手を優先するかを調整できる．しかし，強さに直接影響する部分ではないため DDA-U 単体では動的な強さの調整が困難である．

$$U(s_t, a) = \frac{W(s_t, a)}{N(s_t)} + C\frac{\sqrt{2\ln(N(S_t)+1)}}{(n(S_t, a)+1)}$$

$W(s_t, a)$ は累計価値，$N(s_t)$ はノード s_t の訪問回数，$n(s_t, a)$ は辺 (s_t, a) の訪問回数，C は探索率を表す．

2.4 DDA-S プレイヤー

DDA-S は，Softmax 関数で使用する温度パラメータを変更することで DDA を実現する手法を用いている．T が与えられたとき，次の確率で手を選択する．なお，N_i は MCTS における i 手目のシミュレーション回数である．

$$\frac{N_i^T}{\sum_j N_j^T}$$

DDA-S は温度が高い場合に，MCTS のシミュレーション回数が多い手（強い手）を選択し，低い場合には MCTS のシミュレーション回数が少ない手を選択する傾向にある．なお，$T = 0$ の場合ではランダムな手を選択する．

第 3 章

AlphaDDA の再評価

先行研究 [1] では，AlphaDDA プレイヤーと他の AI プレイヤーを対戦（先手後手入れ替えて 100 試合）させ，その際の勝率をもとに DDA の評価を行なっていた．しかし，勝率のみによる評価では最終的な勝率が 50%だとしても対局途中では AlphaDDA プレイヤーが優勢（または劣勢）の状況が考えられる．

そこで，先行研究で得られた学習結果のパラメータを用い，対局中の局面評価値の変動の大きさの再評価を行った．

3.1　実験設定

対象のゲームは先行研究と同様にリバーシ（オセロ）を使用した．AlphaDDA プレイヤーとしては，先行研究 [1] において使用されたパラメータをそのまま用い，DDA-M，DDA-D，DDA-U プレイヤーを対象に評価実験を行った．

対戦相手プレイヤーとしては以下のプレイヤーを使用した．

- MCTS1 プレイヤー

  モンテカルロ木探索ベースのシミュレーション回数 1 手あたり 300 回のプレイヤー
- MCTS2 プレイヤー

  モンテカルロ木探索ベースのシミュレーション回数 1 手あたり 100 回のプレイヤー
- Minimax プレイヤー

  ミニマックス探索ベースのプレイヤー
- AlphaZero プレイヤー

動的強さ調整を導入する前の AlphaZero プレイヤー

各相手 AI プレイヤーに対し，先手後手入れ替えて 100 試合を行った棋譜を取得し，各棋譜の各手数において非常に強力な AI プレイヤー Edax[5] を用いて評価値を求めた．Edax の評価値については，石数を表すようスケーリングされている．

3.2　結果と考察

各 AI との対局における勝敗の内訳，対局を通した評価値の平均絶対誤差（序盤 8 手分を除く），30 手目における評価値の結果を表 3.1〜3.3 に示す.

MCTS1，MCTS2，AlphaZero，Minimax プレイヤーそれぞれとの局面評価を図 3.1〜3.12 に示す.

3.2.1　DDA-M の結果

表 3.1　DDA-M の対局結果

相手	勝敗			平均絶対誤差	30 手目
	勝	負	分	（平均 ±SD）	（平均 ±SD）
Minimax	62	35	3	16.3 ± 8.3	11.5 ± 19.4
MCTS1	42	50	8	19.6 ± 10.1	19.5 ± 25.2
MCTS2	59	36	5	22.9 ± 11.7	22.2 ± 28.5
AlphaZero	58	41	1	14.4 ± 5.8	−4.8 ± 17.7

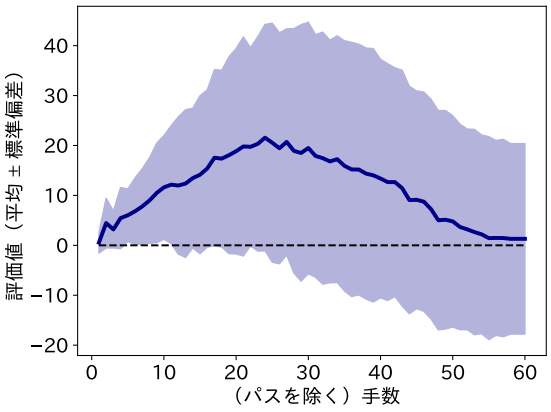

図 3.1　DDA-M vs MCTS1 の局面評価

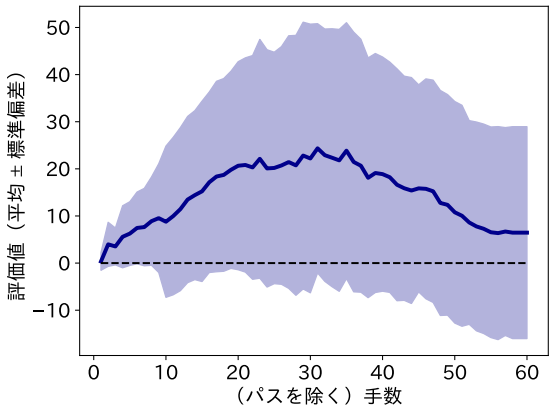

図 3.2　DDA-M vs MCTS2 の局面評価

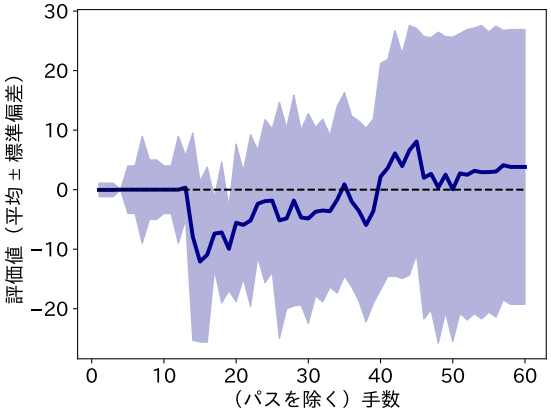

図 3.3　DDA-M vs AlphaZero の局面評価

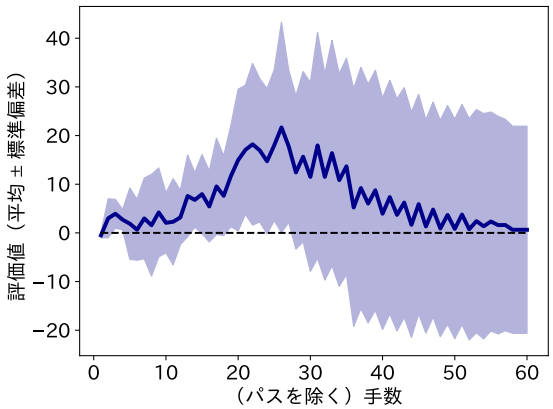

図 3.4　DDA-M vs Minimax の局面評価

3.2.2　DDA-D の結果

表 3.2　DDA-D の対局結果

相手	勝敗			平均絶対誤差	30 手目
	勝	負	分	（平均 ±SD）	（平均 ±SD）
Minimax	72	23	5	16.0 ± 20.9	17.2 ± 20.9
MCTS1	43	50	7	18.6 ± 9.0	15.9 ± 25.3
MCTS2	54	40	6	19.3 ± 7.8	20.7 ± 20.2
AlphaZero	43	52	5	11.9 ± 4.3	2.0 ± 12.7

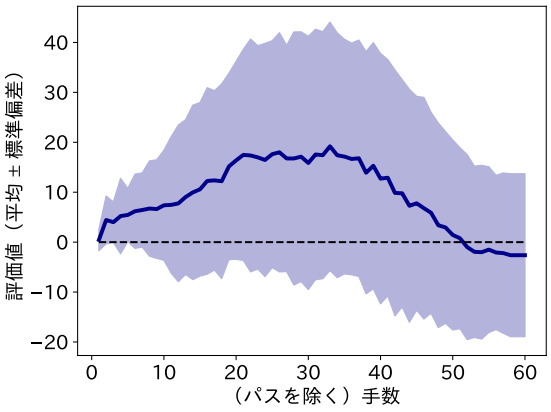

図 3.5　DDA-D vs MCTS1 の局面評価

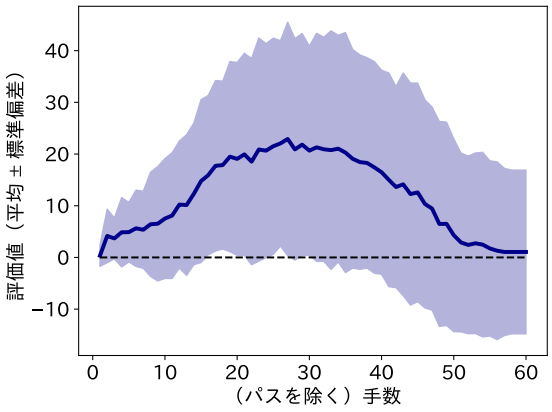

図 3.6　DDA-D vs MCTS2 の局面評価

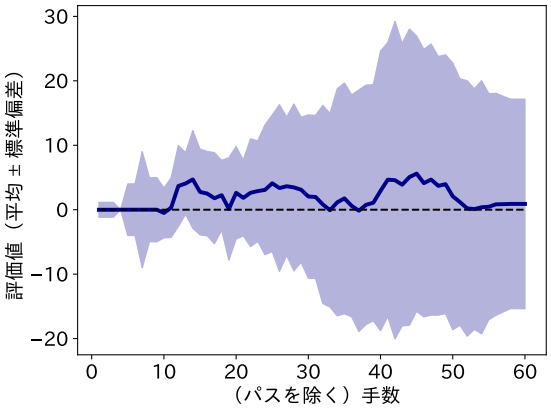

図 3.7　DDA-D vs AlphaZero の局面評価

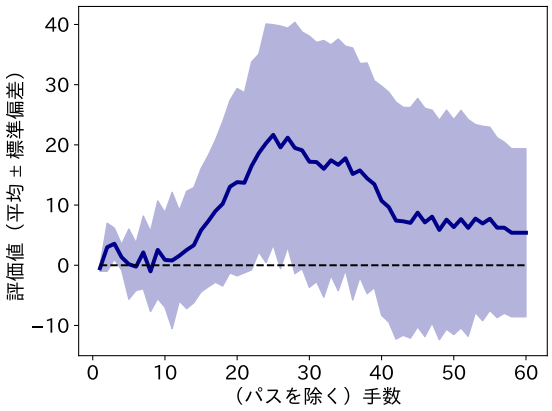

図 3.8　DDA-D vs Minimax の局面評価

3.2.3　DDA-U の結果

表 3.3　DDA-U の対局結果

相手	勝敗			平均絶対誤差	30 手目
	勝	負	分	（平均 ±SD）	（平均 ±SD）
Minimax	13	71	16	12.9 ± 2.9	−1.8 ± 10.7
MCTS1	1	99	0	20.5 ± 7.3	−16.0 ± 19.1
MCTS2	3	93	4	18.2 ± 6.3	−14.9 ± 20.8
AlphaZero	0	100	0	30.5 ± 6.5	−35.7 ± 11.3

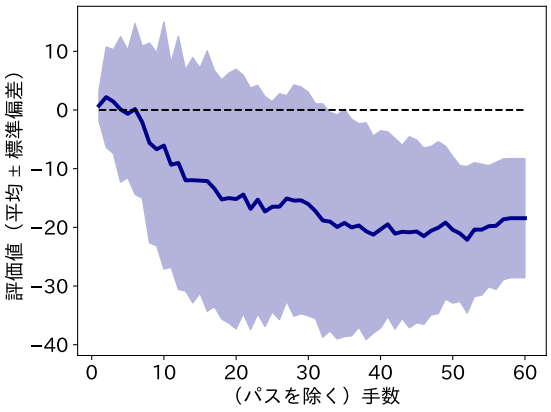

図 3.9　DDA-U vs MCTS1 の局面評価

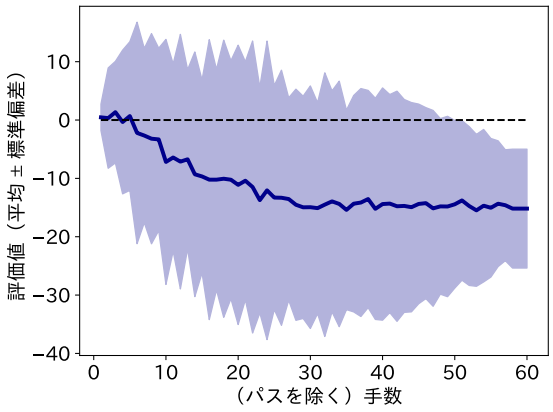

図 3.10　DDA-U vs MCTS2 の局面評価

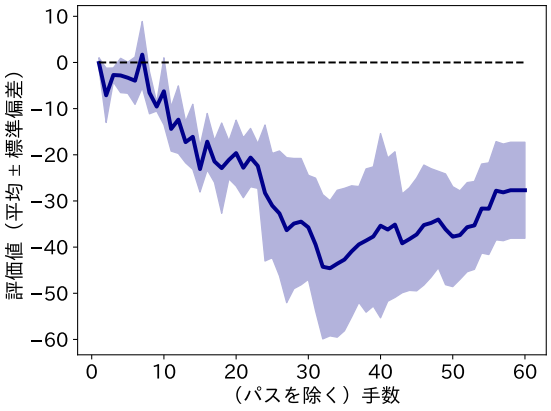

図 3.11　DDA-U vs AlphaZero の局面評価

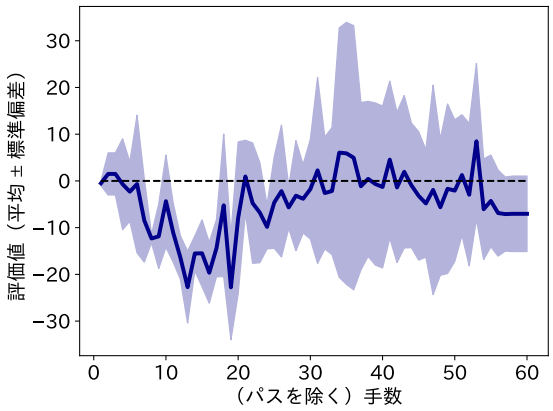

図 3.12　DDA-U vs Minimax の局面評価

3.2 結果と考察

勝率に関しては，先行研究 [1] と同等の結果が得られた．しかしながら，対局を通した評価値の平均絶対誤差は 14.4〜22.9 であり，局面の優劣の中立を保てていないことが分かる．ゲームの中盤までに評価値が大きく振れてしまい，その評価値の幅はゲーム終盤でもほとんど小さくなっていない．

DDA-M の局面評価では，対 MCTS1・MCTS2・Minimax では 25〜30 手目までの局面では DDA-M プレイヤーは相手よりも強い手を打ち，終局に近づくにつれてより弱い手を打っている．対 AlphaZero では評価値は上下するものも，全体を通してどちらかのプレイヤーに有利な局面が継続する訳ではなかった．

DDA-D の局面評価では，DDA-M と同様に対 MCTS1・MCTS2・Minimax では DDA-D が局面中盤で弱い手を打つように調整している．対 AlphaZero では，ゲーム全体を通して DDA-D にとって有利な局面が続くが，平均の評価値の揺れ幅は他の AI と対局させた場合より小さい．

DDA-U はそもそも勝率 50%への調整も行われなかった．局面評価では，対 MCTS1・MCTS2・AlphaZero において 30 手目までに評価値が下がり続けていた．

これらの結果から，AlphaDDA は DDA-M と DDA-D において勝率の点では目標を達成できているものの，ゲーム全体に渡って適切に強さを制御できているとは言えない．

第 4 章

合議

強いゲーム AI プレイヤーを作成する際に，合議システムと呼ばれる手法の研究が行われてきた．合議システムは，複数の思考プログラムの候補手から一つの手を選択する手法である．将棋における合議では，単純な多数決や楽観合議と呼ばれる手法において単一思考のプレイヤーよりも強いプレイヤーが作成可能であることが示されている [3].

強いプレイヤーを作る上では，多数決による合議では，最も多くのプレイヤーが支持した手を選択する．楽観合議と呼ばれる評価値の最も高い手を選択する合議もある．

本研究では，合議により複数の候補手からより互角に近づく手を選択するプレイヤーを作成し，DDA プレイヤーの動的な強さの調整の改良を目指す.

4.1 DDA プレイヤーの合議

合議により AlphaDDA の強さの調整の改良を行う.

本研究では，強さの調整を実現する合議の手法として多数決合議と楽観合議を使用する.

多数決合議では，強いプレイヤーを作成する時と同様に複数の候補手の中で最も多くのプレイヤーが支持した手を選択する．楽観合議では，候補手の中から AlphaZero により求められた評価値が最も 0 に近づく手を選択するものとする.

また，各 DDA プレイヤーが同じ手しか支持しない場合，合議による複数の手から最も良い手を選択することができない．そこで，DDA プレイヤーの強さの調整と乱数による複数プレイヤーの生成を行なった．乱数は，各 DDA プレイヤーが手を選択する際の評価値に加えている.

まず，DDA-U を除く各 DDA プレイヤーで合議を行った際により多くの候補手が得られるように，評価値における強さの調整を行なった．

4.1.1　DDA-M の調整

評価値とシミュレーション回数の関係を 4.1 に示す．$player$ は，先手プレイヤーが 1 後手プレイヤーが -1 である．

評価値が -1 以下ではシミュレーション回数は最大の 400 にした．また，最小のシミュレーション回数は 1 である．

$$N_{sim}(\bar{v}_n) = \lceil 120^{-1.2528\bar{v}_n * player} - 20 \rceil$$

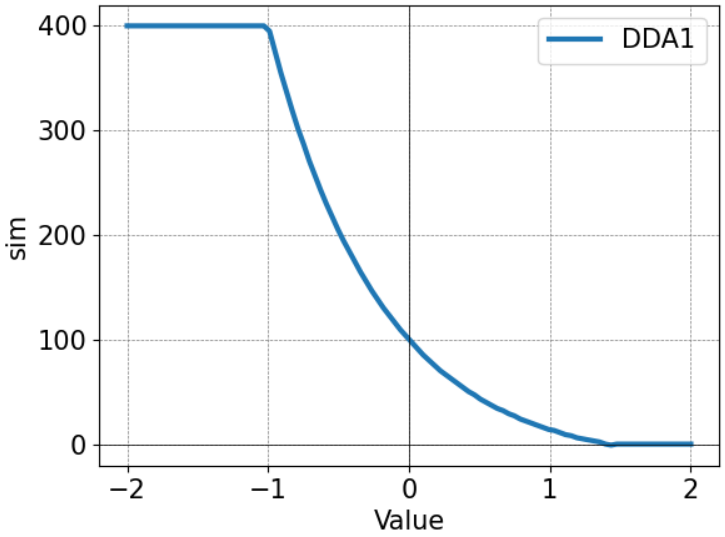

図 4.1　DDA-M の評価値とシミュレーション回数

4.1.2　DDA-D の調整

評価値とドロップアウト確率の関係を図 4.2 に示す．評価値 0 以下ではドロップアウト確率は 0 とし，最大のドロップアウトは 0.95 とした．

$$P_{drop(\bar{V}_n)} = V_n * player$$

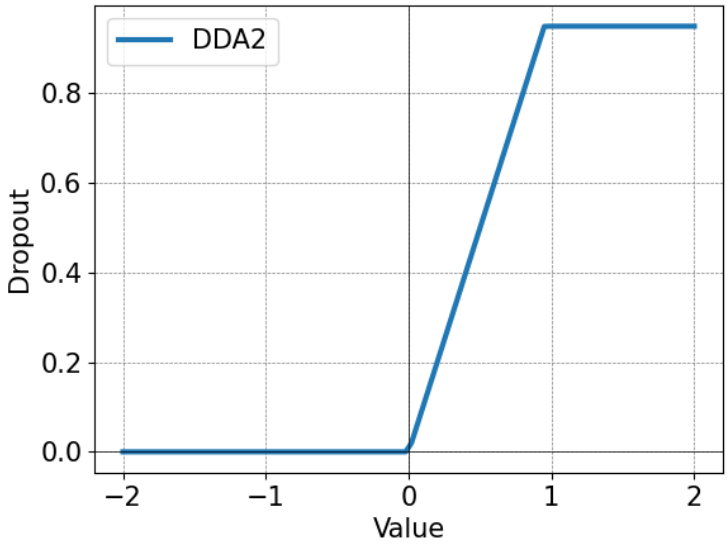

図 4.2　DDA-D の評価値とドロップアウト確率

4.1.3　DDA-U の調整

本研究の合議実験において C の調整は特に行わず，先行研究のパラメータ $C = 0.6$ をそのまま使用した．

4.1.4　DDA-S の調整

評価値と温度パラメータの関係を図 4.3 に示す．また，単一の DDA-S の対局結果を表 4.1 に，その時の局面評価を図 4.4〜図 4.7 に示す．

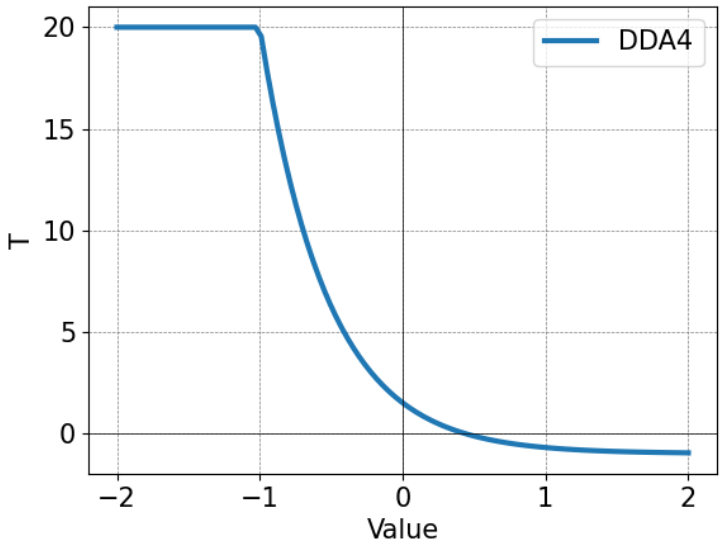

図 4.3　DDA-S の評価値と温度パラメータ

表 4.1　DDA-S の対局結果

相手	勝敗			平均絶対誤差	30 手目
	勝	負	分	（平均 ±SD）	（平均 ±SD）
Minimax	47	50	3	15.8 ± 9.9	0.5 ± 22.4
MCTS1	13	82	5	15.3 ± 5.2	−3.8 ± 20.0
MCTS2	18	77	5	14.4 ± 6.1	0.7 ± 20.4
AlphaZero	12	86	2	15.4 ± 6.9	−11.0 ± 16.2

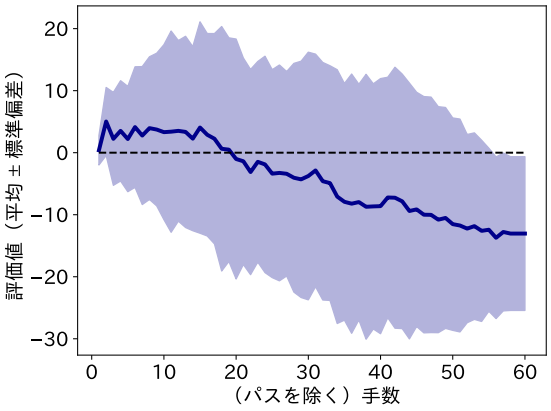

図 4.4　DDA-S vs MCTS1 の局面評価

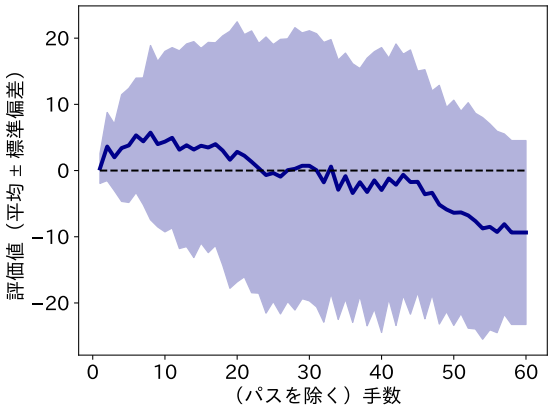

図 4.5　DDA-S vs MCTS2 の局面評価

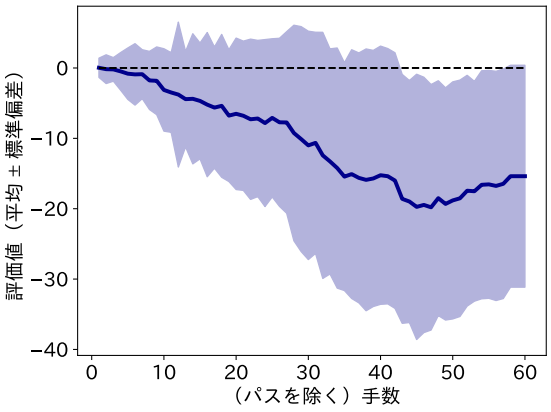

図 4.6　DDA-S vs AlphaZero の局面評価

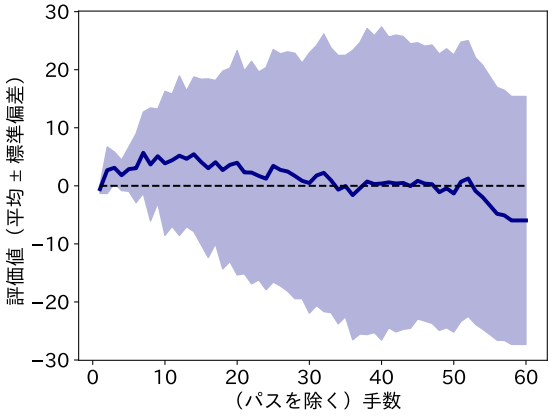

図 4.7　DDA-S vs Minimax の局面評価

4.2　予備実験

合議に向けて，前節で調整を行なった DDA プレイヤーに対して，乱数による複数プレイヤーの生成を行う前に，手を選択する際の評価値に幾つかの値を加えた際の候補手の変化の調査を行なった．

加えた値は $-1, -0.8, -0.6, -0.4, -0.2, -0.1, -0.05, 0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0$ である．

その際，オセロの棋譜を 500 試合分用意し，それらの棋譜の 19 手置いた後の局面について Edax の評価値をもとに有利・不利の局面に分割した．後手を DDA プレイヤーとして，先手から見た評価値をもとに以下の 3 つに分けた．

- 勝ち局面（大）：評価値 -64〜-18 の 154 局
- 勝ち局面（小）：評価値 -17〜-1 の 155 局
- 負け局面：評価値 0〜64 の 191 局

それぞれの局面において加えた値と手が変化したかを調べた．各 DDA プレイヤーが局面の状況に対して，AlphaZero による評価値から求めた強い手上位 3 つを選択した割合を図 4.8〜図 4.11 に示す．

負け局面では強い手を選択し，勝ち局面では強い手を選択する割合が少ないことがわかる．後手プレイヤーにおいて，評価値にマイナスの値を加えるほど自身に有利な局面であると判断することになるため上位の手を選ぶ割合が減ることがわかる．

値を加えることで手の強さに変化が起きていることが確認できた．

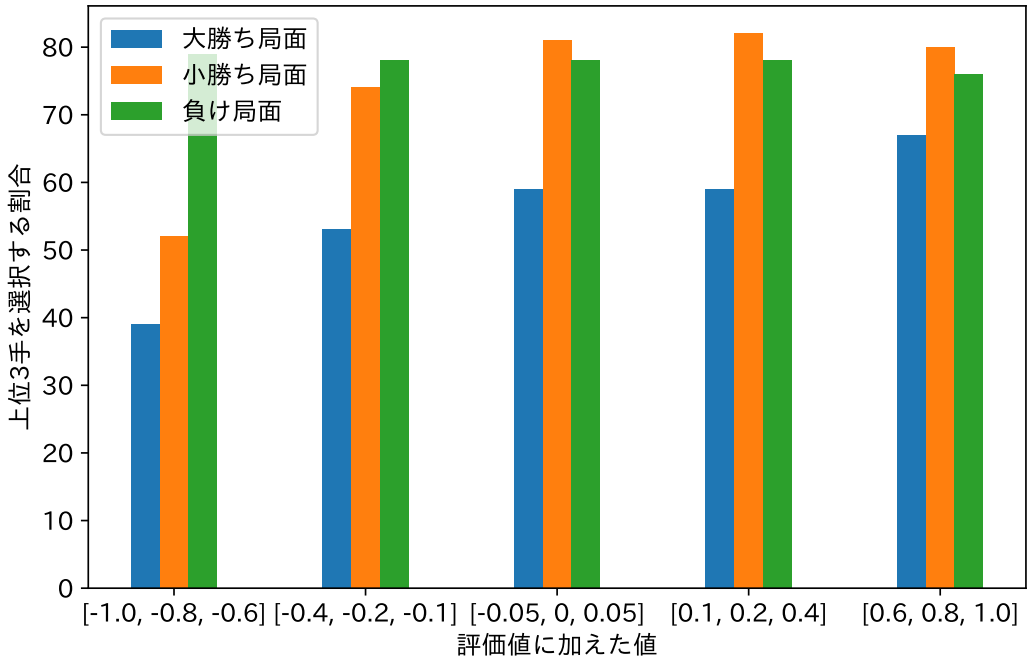

図 4.8　DDA-M の選択した手

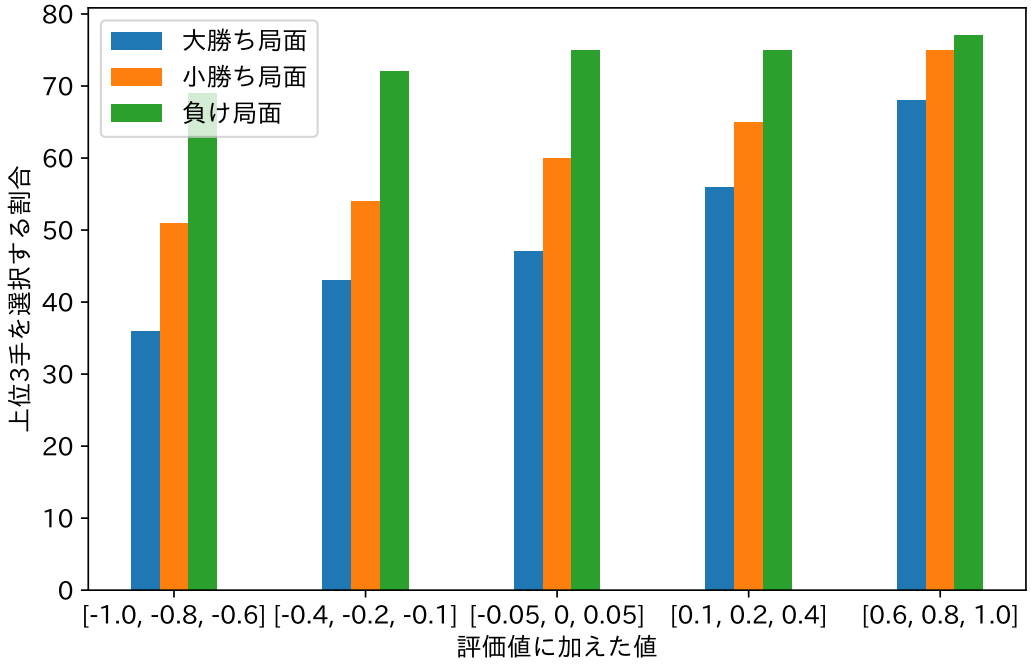

図 4.9　DDA-D の選択した手

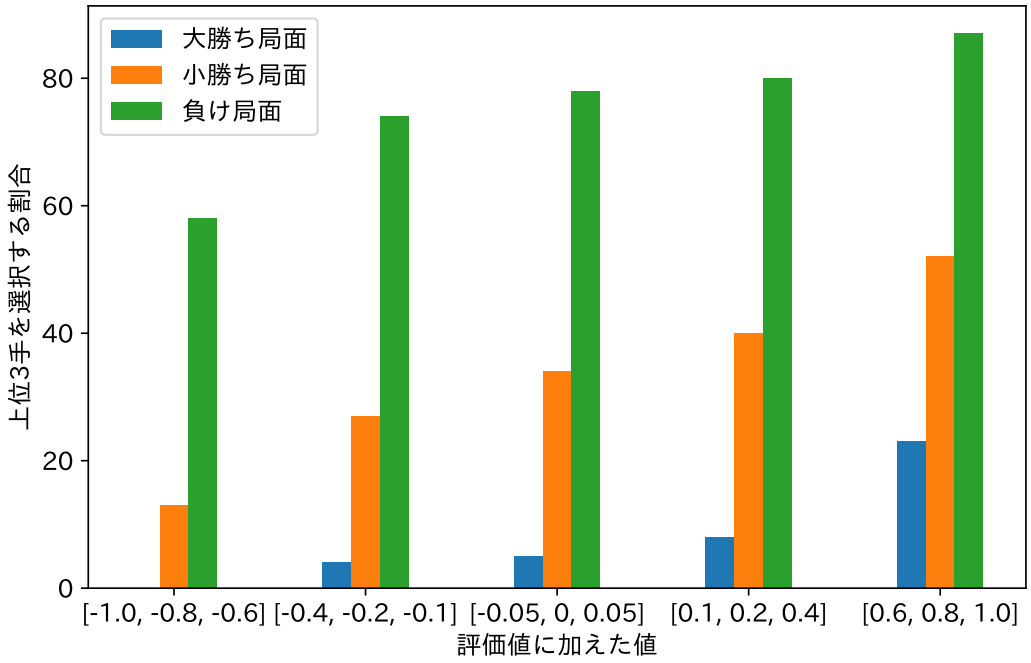

図 4.10　DDA-U の選択した手

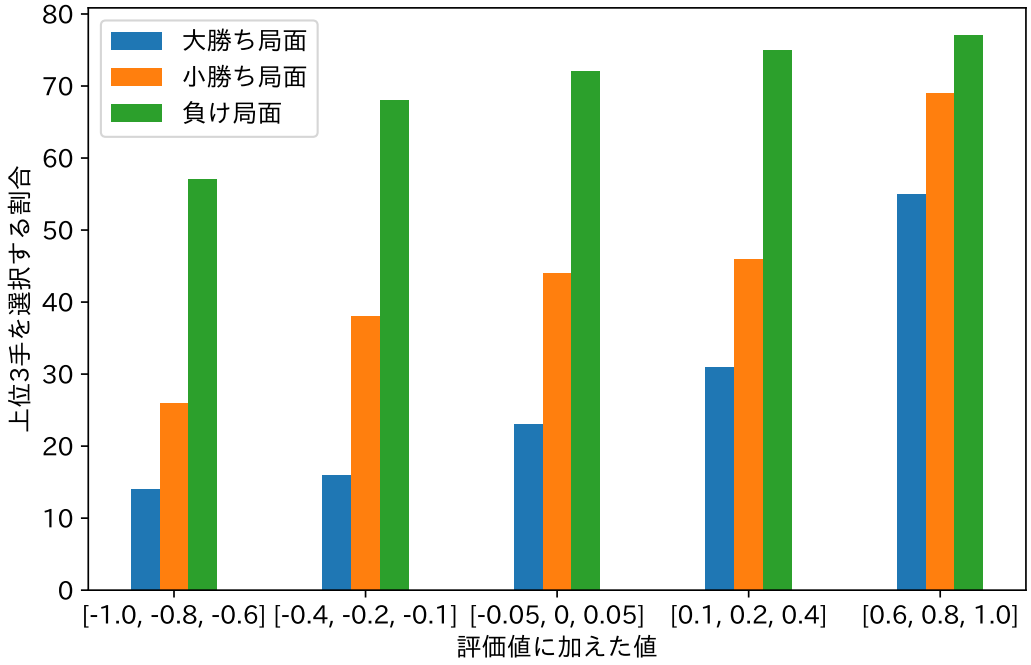

図 4.11　DDA-S の選択した手

第 5 章

実験

伊藤らによる Bonanza の多数決合議実験では，プレイヤー数が増えると合議プレイヤーがより強くなる傾向があった [3]．本研究では伊藤らの実験で使われた最大のプレイヤー数である 16 プレイヤーを使用した場合の合議実験の結果を示す．

多数決合議プレイヤーと楽観合議プレイヤーそれぞれにおいて，各 AI と先手後手入れ替えて計 100 試合行い，勝率及び局面評価値を調査する．合議プレイヤーには乱数として標準偏差 0.2, 0.6, 1.0 のうちのどれかを各 DDA プレイヤーに 4 つずつ加え，16 プレイヤーでの合議を行った場合の結果を掲載する．

5.1 多数決合議の実験結果

表 5.1 に多数決合議プレイヤーとの対局結果を示す．

対局結果から，勝率に関しては MCTS1 や MCTS2 を相手にした時に単一の DDA-M・DDA-D よりも下がっていること，加える乱数の標準偏差が大きい方が勝率が高いという結果である．

これは，合議のために多くの手の選択肢を作ろうとした段階で先行研究のパラメータのプレイヤーより弱いプレイヤーになってしまったことが考えられる．

表 5.1　多数決合議プレイヤーとの対局結果

相手 AI	加えた乱数の標準偏差と勝敗								
	0.2			0.6			1		
	勝	負	分	勝	負	分	勝	負	分
MCTS1	28	64	8	23	70	7	35	57	8
MCTS2	32	62	6	39	52	9	48	43	9
Minimax	63	34	3	66	22	12	66	28	6
AlphaZero	11	83	6	15	81	4	22	73	5

5.1.1　多数決合議プレイヤー vs MCTS1

多数決合議プレイヤーと MCTS1 の局面評価の結果を図 5.1〜図 5.3 に示す.

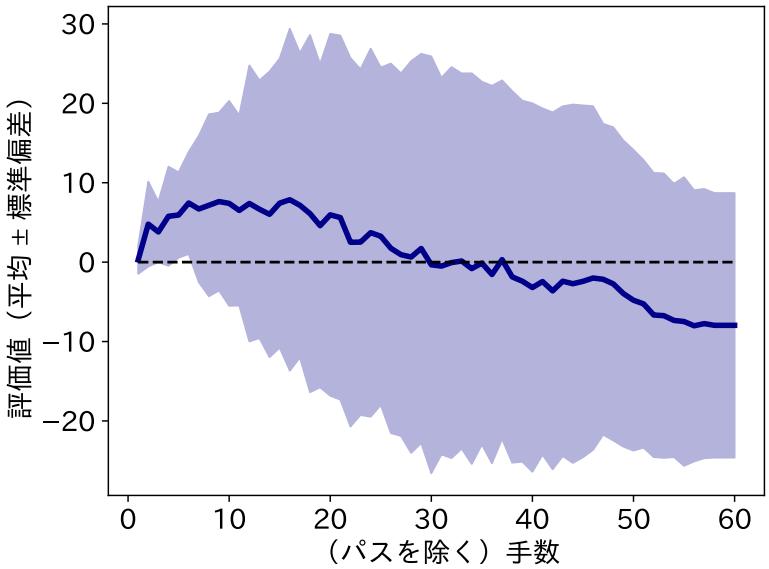

図 5.1　標準偏差 0.2 の乱数を加えた多数決合議プレイヤー vs MCTS1 の局面評価

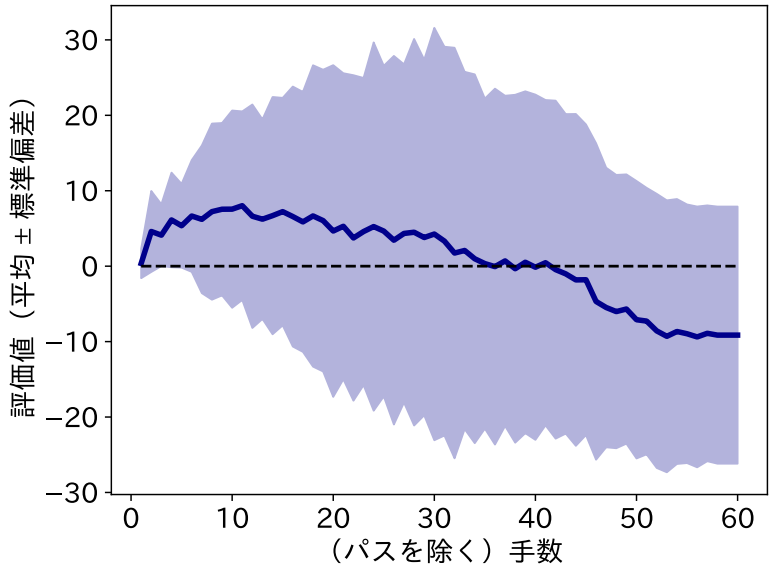

図 5.2　標準偏差 0.6 の乱数を加えた多数決合議プレイヤー vs MCTS1 の局面評価

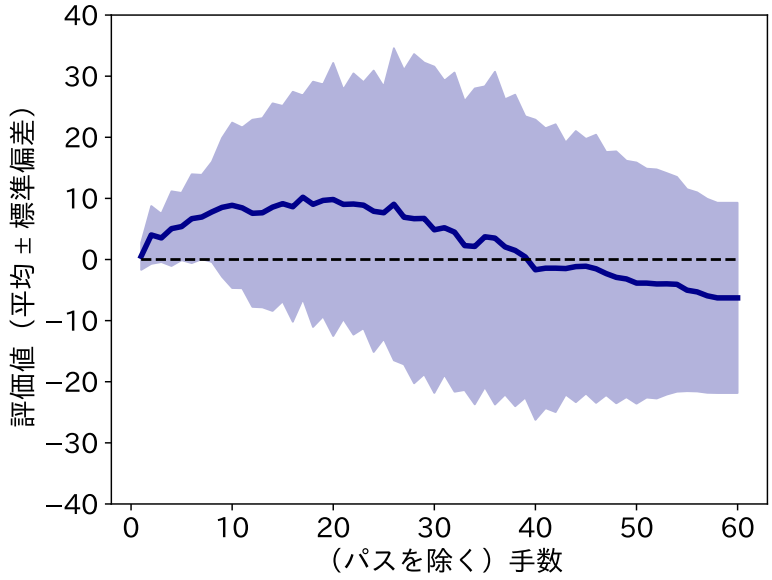

図 5.3　標準偏差 1.0 の乱数を加えた多数決合議プレイヤー vs MCTS1 の局面評価

MCTS1 vs AlphaDDA の再評価の際にはどの DDA でも 30 手目ほどで評価値に 20 ほどの差が見られていたが，多数決合議プレイヤーでは平均で ±10 ほどに抑えられた．また，合議プレイヤーでは 40 手目ほどで有利から不利に変わるという結果になった．MCTS1 に対しては，乱数による評価値の大きな違いは見られなかった．

5.1.2　多数決合議プレイヤー vs MCTS2

多数決合議プレイヤーと MCTS2 の局面評価の結果を図 5.4〜図 5.6 に示す．

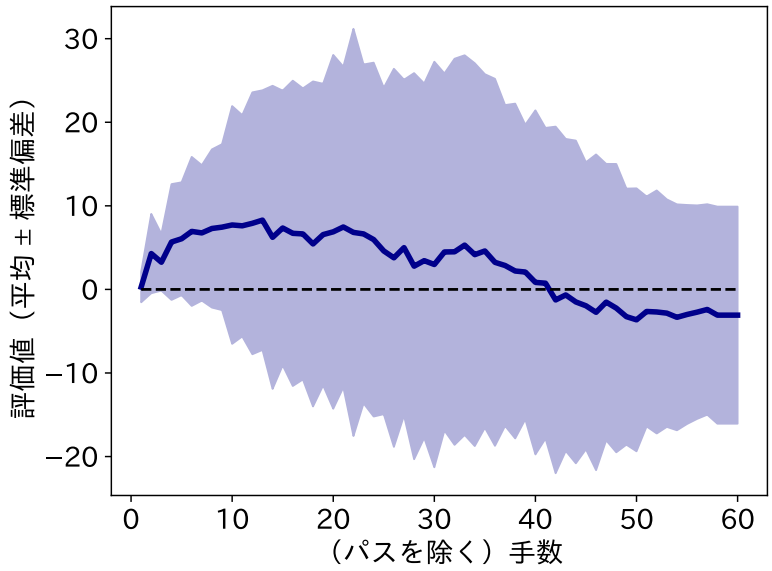

図 5.4　標準偏差 0.2 の乱数を加えた多数決合議プレイヤー vs MCTS2 の局面評価

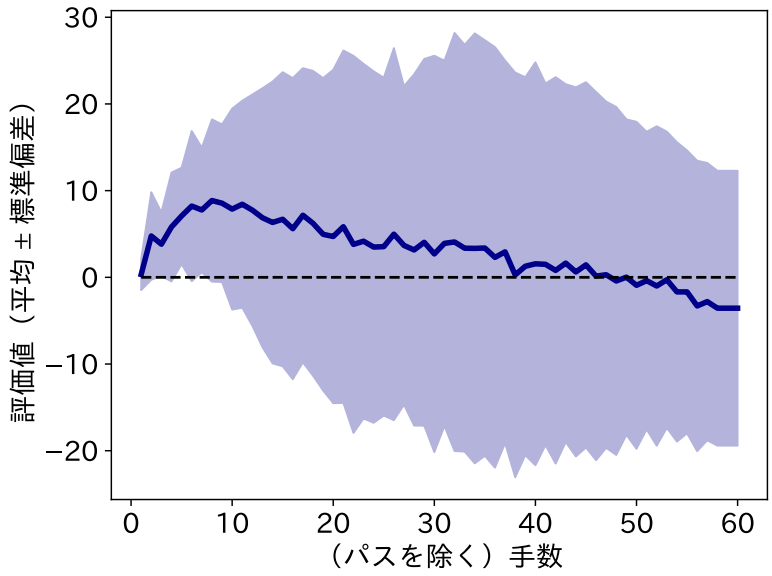

図 5.5　標準偏差 0.6 の乱数を加えた多数決合議プレイヤー vs MCTS2 の局面評価

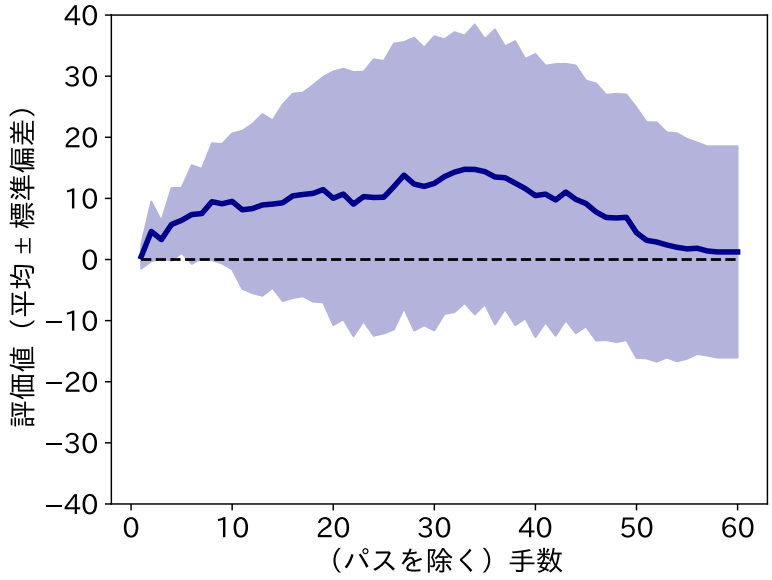

図 5.6　標準偏差 1.0 の乱数を加えた多数決合議プレイヤー vs MCTS2 の局面評価

多数決合議プレイヤーと MCTS2 の局面評価では，単一の DDA では 30 手目付近で 20 ほどであった評価値が抑えられたことがわかる．乱数の標準偏差が 1.0 の時に 0.2 や 0.6 に比べて評価値がプラスによっており，勝率も高い．

5.1.3　多数決合議プレイヤー vs Minimax

多数決合議プレイヤーと Minimax の局面評価の結果を図 5.7〜図 5.9 に示す．

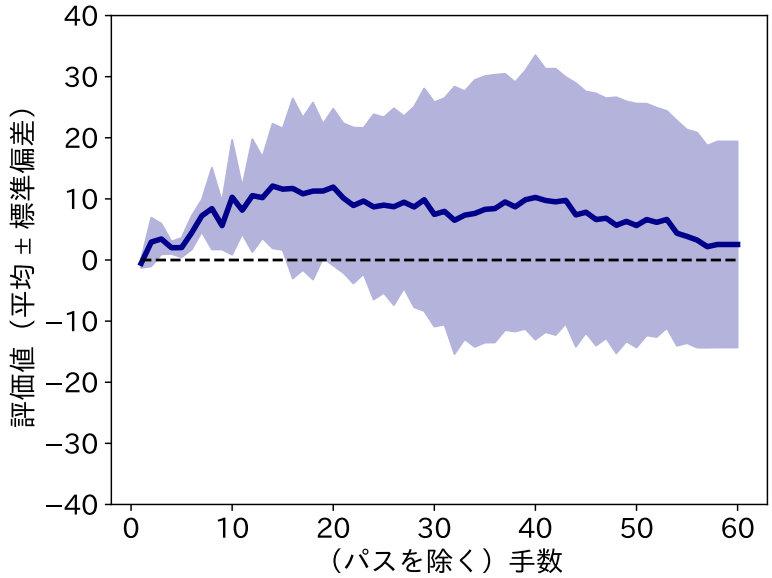

図 5.7　標準偏差 0.2 の乱数を加えた多数決合議プレイヤー vs Minimax の局面評価

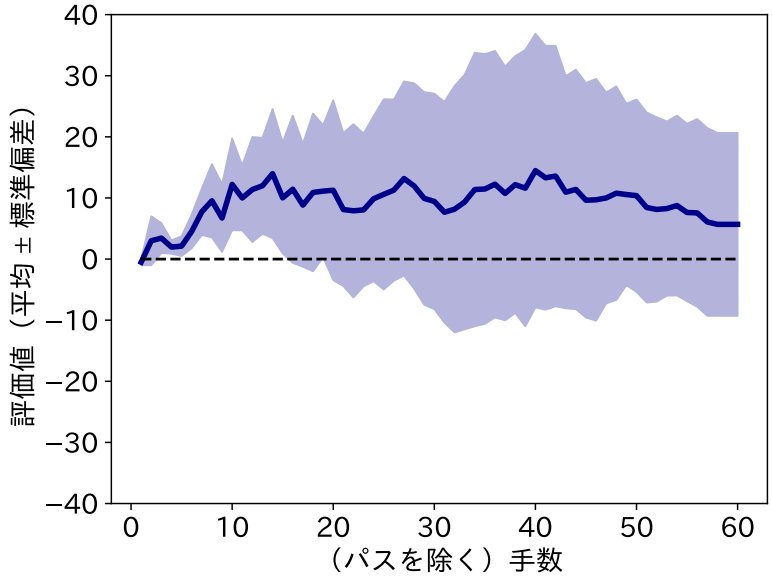

図 5.8　標準偏差 0.6 の乱数を加えた多数決合議プレイヤー vs Minimax の局面評価

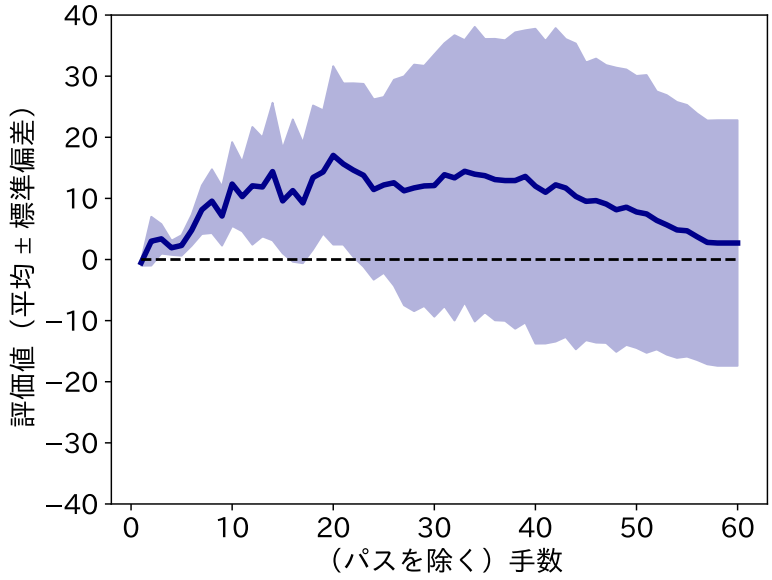

図 5.9　標準偏差 1.0 の乱数を加えた多数決合議プレイヤー vs Minimax の局面評価

5.1.4　多数決合議プレイヤー vs AlphaZero

多数決合議プレイヤーと AlphaZero の局面評価の結果を図 5.10〜図 5.12 に示す.

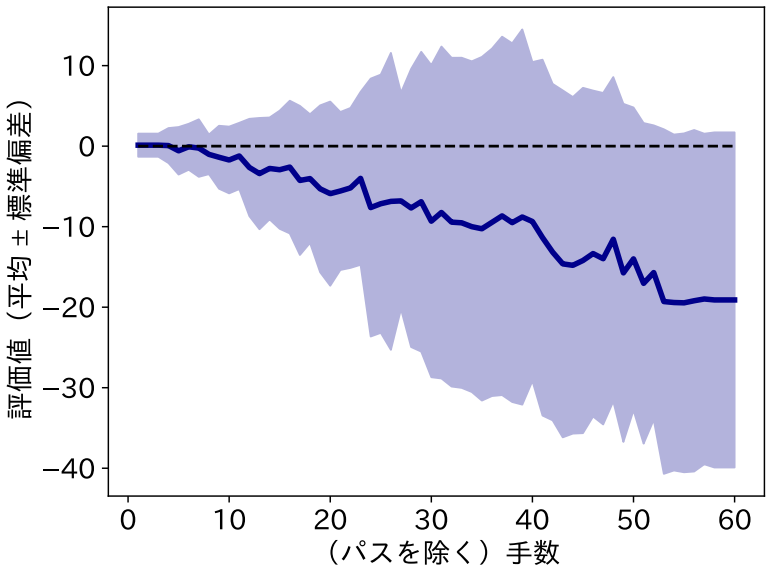

図 5.10　標準偏差 0.2 の乱数を加えた多数決合議プレイヤー vs Alphazero の局面評価

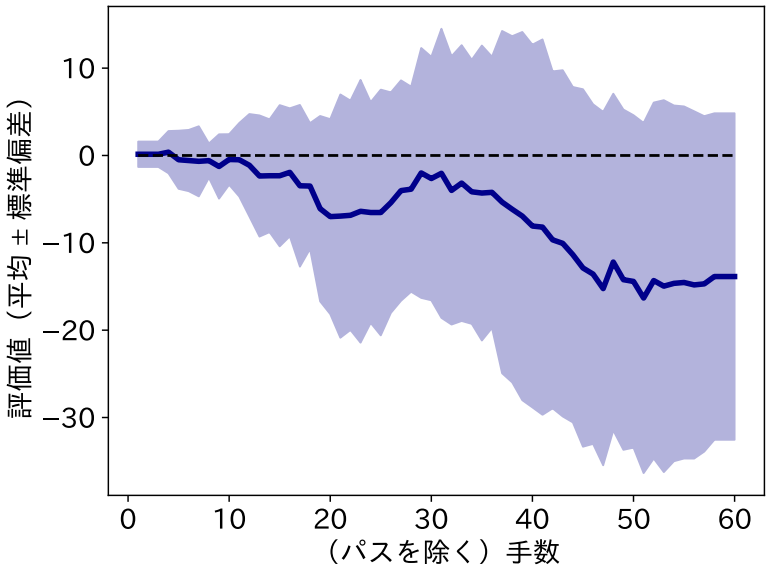

図 5.11　標準偏差 0.6 の乱数を加えた多数決合議プレイヤー vs Alphazero の局面評価

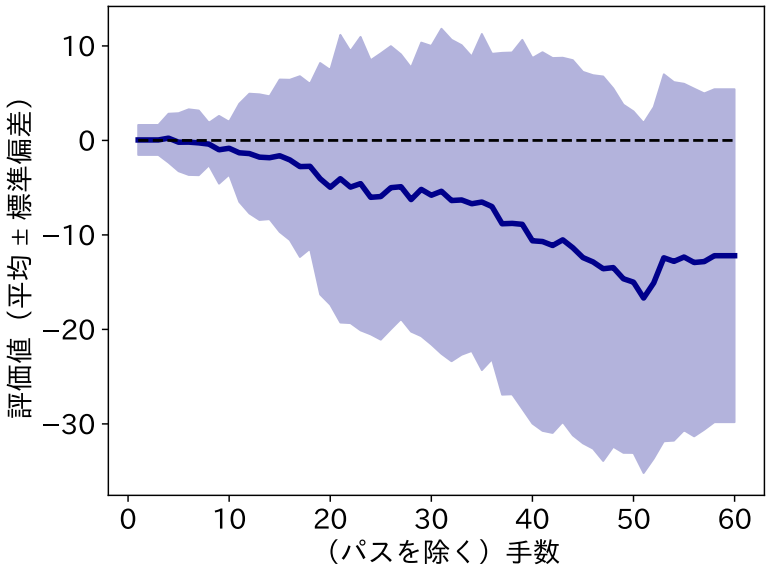

図 5.12　標準偏差 1.0 の乱数を加えた多数決合議プレイヤー vs Alphazero の局面評価

多数決合議プレイヤーと Minimax では，ゲーム全体を通して評価値の平均はプラスであり，勝率としても 6 割以上であった．

勝率に関しては，多数決合議プレイヤーは，MCTS1 と 2 に比べると弱く Minimax プレイヤーよりは強い．局面評価では，単一の DDA プレイヤーと比較すると評価値の変動の幅が小さかった．

5.2　楽観合議の実験結果

表 5.2 に楽観合議プレイヤーとの対局結果を示す．（対 AlphaZero プレイヤーとの結果は実験中のため掲載していない．）

勝率に関しては，MCTS1 と 2 に対しては極端に低く，Minimax 相手にも 50%もない．単一の DDA-U に比べると勝率は高いがそれ以外のプレイヤーと比べると勝率は低い．

一方，引き分けは多く評価値 0 を目指そうとしていることはわかる．

表 5.2　楽観合議プレイヤーとの対局結果

相手 AI	加えた乱数の標準偏差と勝敗								
	0.2			0.6			1		
	勝	負	分	勝	負	分	勝	負	分
MCTS1	3	85	12	4	81	15	2	83	15
MCTS2	4	77	19	6	79	15	4	82	14
Minimax	41	34	25	25	51	24	31	37	32
Alphazero	8	76	16	9	78	13	12	71	17

5.2.1　楽観合議プレイヤー vs MCTS1

楽観合議プレイヤーと MCTS1 の局面評価の結果を図 5.13〜図 5.15 に示す.

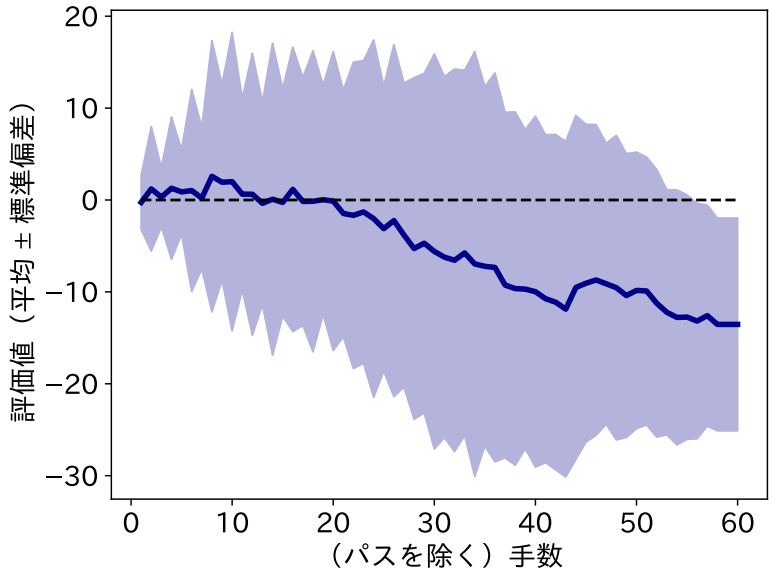

図 5.13　標準偏差 0.2 の乱数を加えた楽観合議プレイヤー vs MCTS1 の局面評価

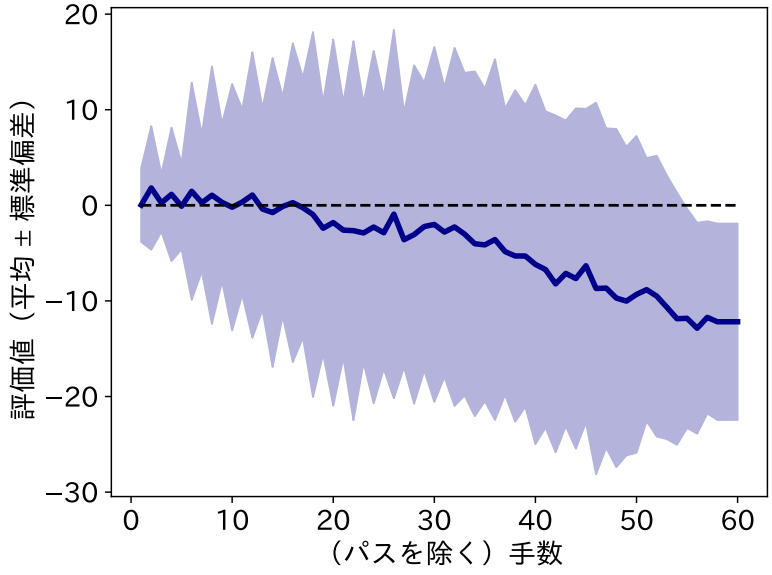

図 5.14　標準偏差 0.6 の乱数を加えた楽観合議プレイヤー vs MCTS1 の局面評価

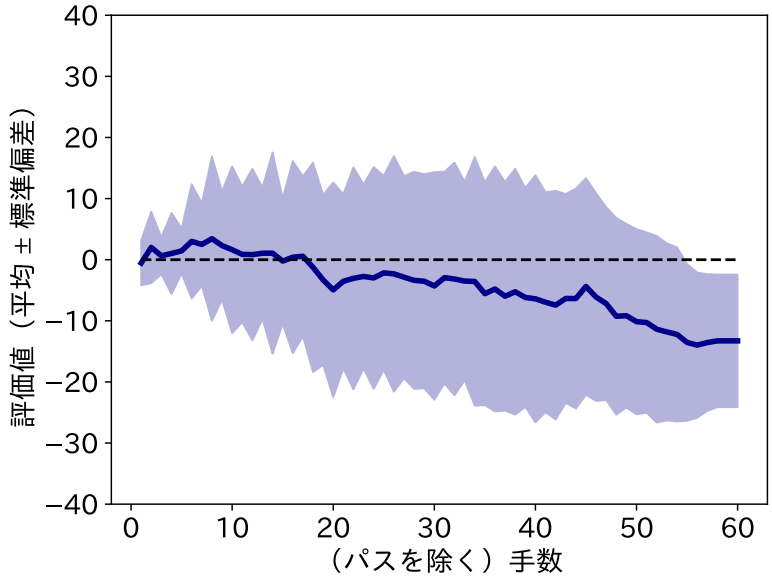

図 5.15　標準偏差 1.0 の乱数を加えた楽観合議プレイヤー vs MCTS1 の局面評価

楽観合議プレイヤーと MCTS1 の局面評価は，20 手目辺りまで評価値 0 に近い状態が続いていたが，その後マイナスに振れる一方である．乱数の大きさによる大きな違いは見られなかった．

5.2.2　楽観合議プレイヤー vs MCTS2

楽観合議プレイヤーと MCTS2 の局面評価の結果を図 5.16〜図 5.18 に示す．楽観合議プレイヤーと MCTS2 プレイヤーでも，20 手目を境に評価値はマイナスに下がる一方である．与えた乱数による大きな違いも見られない．

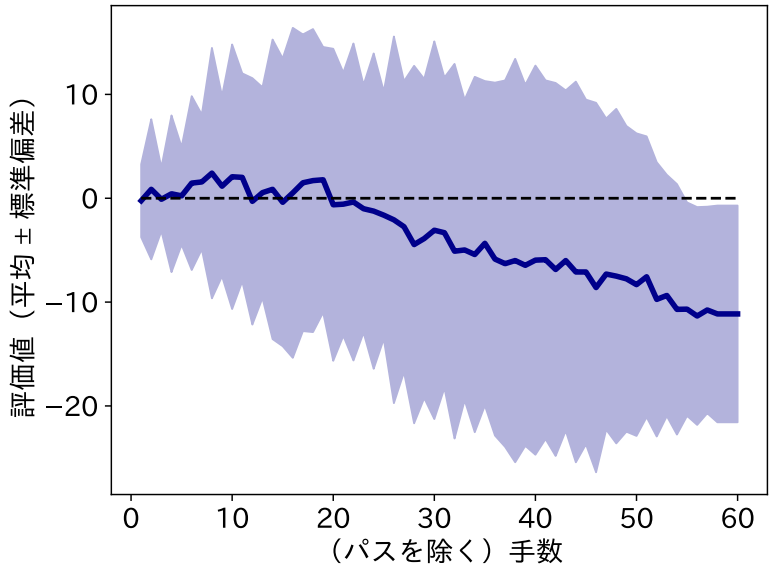

図 5.16　標準偏差 0.2 の乱数を加えた楽観合議プレイヤー vs MCTS2 の局面評価

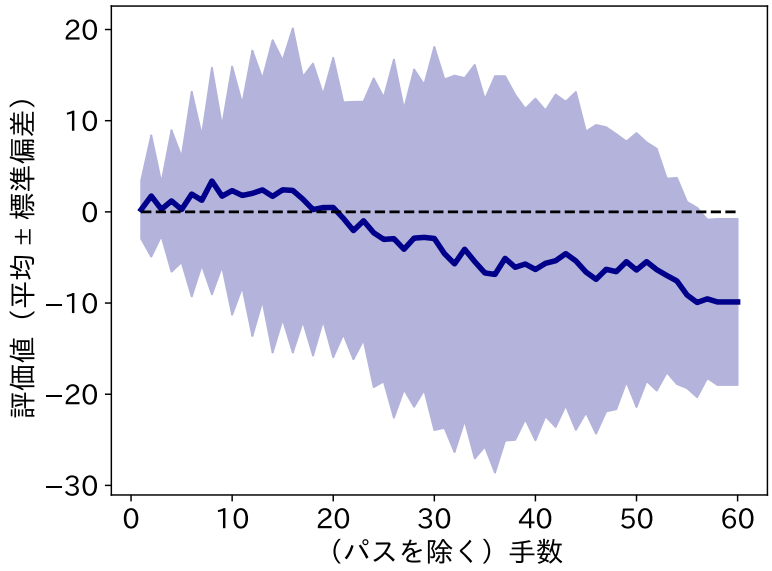

図 5.17　標準偏差 0.6 の乱数を加えた楽観合議プレイヤー vs MCTS2 の局面評価

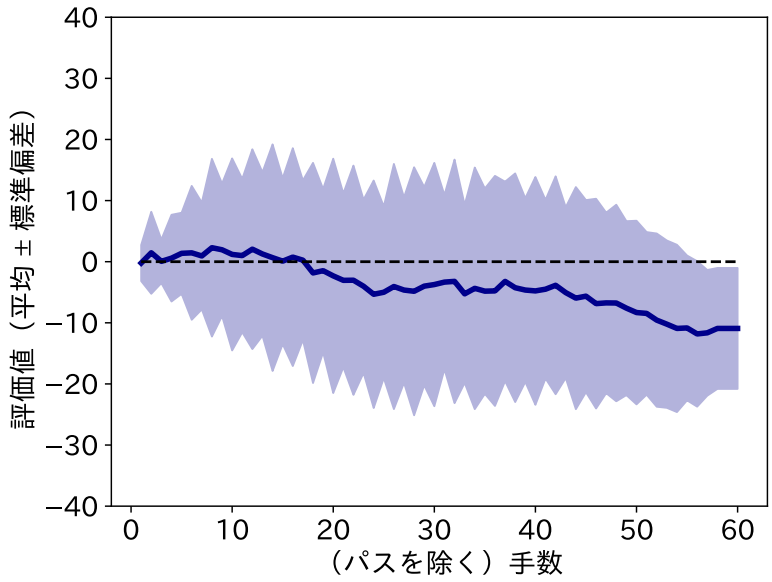

図 5.18　標準偏差 1.0 の乱数を加えた楽観合議プレイヤー vs MCTS2 の局面評価

5.2.3　楽観合議プレイヤー vs Minimax

楽観合議プレイヤーと Minimax の局面評価の結果を図 5.19〜図 5.21 に示す．楽観合議プレイヤーと Minimax では，評価値の平均は ±10 の範囲で変動していた．どの乱数の局面評価でも 15 手目辺りで評価値がプラス側に上がり，55 手目辺りでマイナスに下がるような形であった．

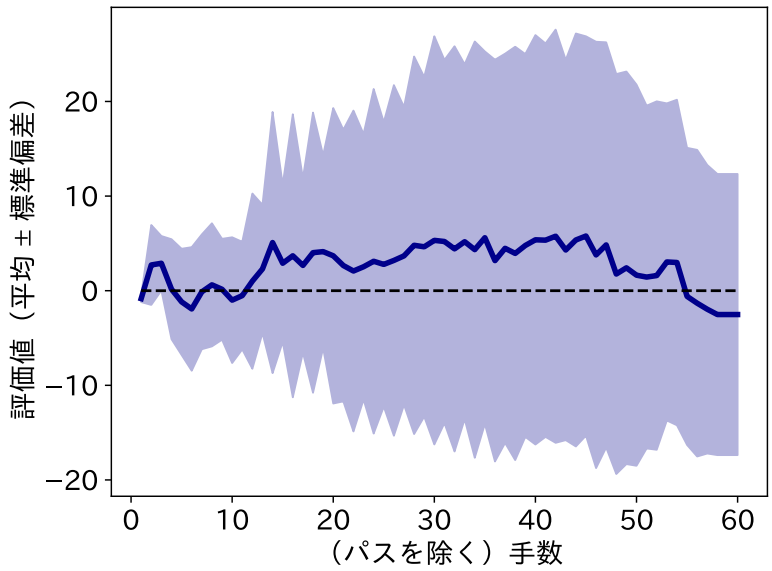

図 5.19　標準偏差 0.2 の乱数を加えた楽観合議プレイヤー vs Minimax の局面評価

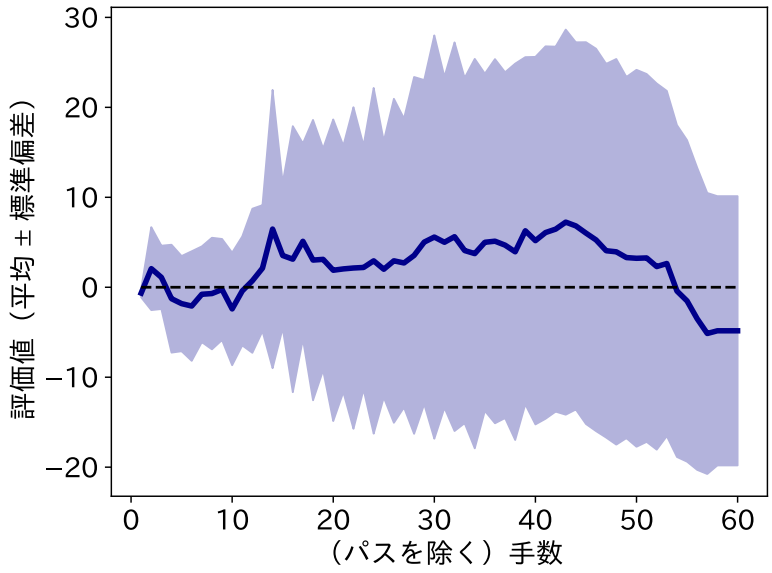

図 5.20　標準偏差 0.6 の乱数を加えた楽観合議プレイヤー vs Minimax の局面評価

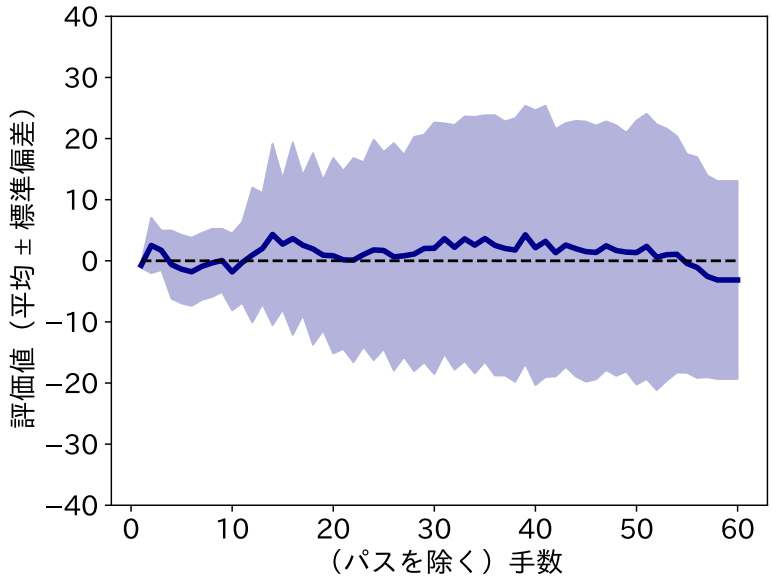

図 5.21　標準偏差 1.0 の乱数を加えた楽観合議プレイヤー vs Minimax の局面評価

5.2.4　楽観合議プレイヤー vs AlphaZero

楽観合議プレイヤーと AlphaZero の局面評価の結果を図 5.22〜図 5.24 に示す.

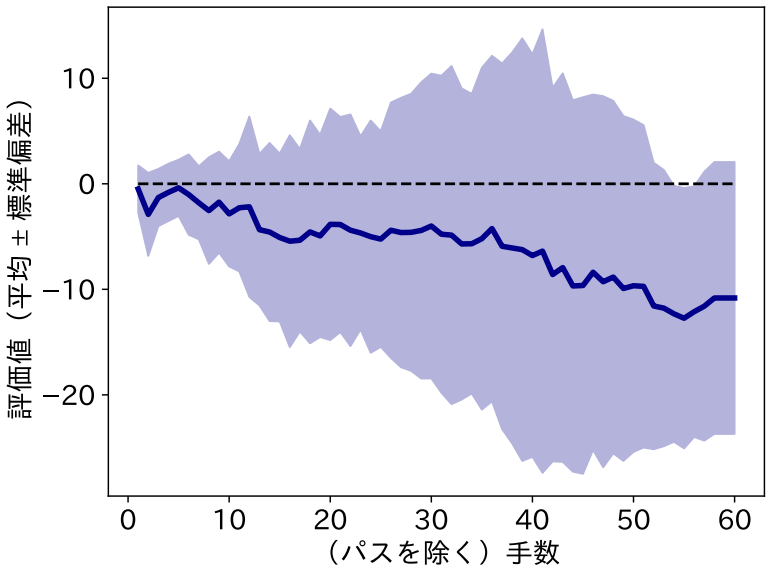

図 5.22　標準偏差 0.2 の乱数を加えた楽観合議プレイヤー vs AlphaZero の局面評価

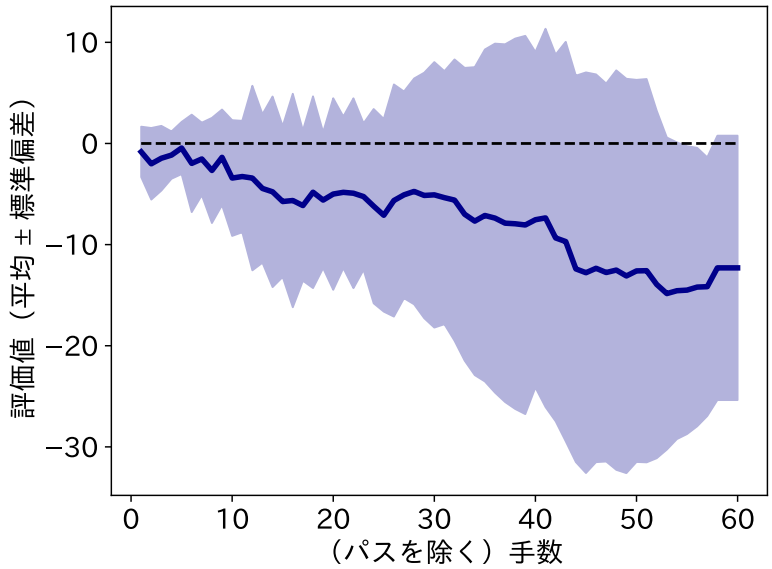

図 5.23　標準偏差 0.6 の乱数を加えた楽観合議プレイヤー vs AlphaZero の局面評価

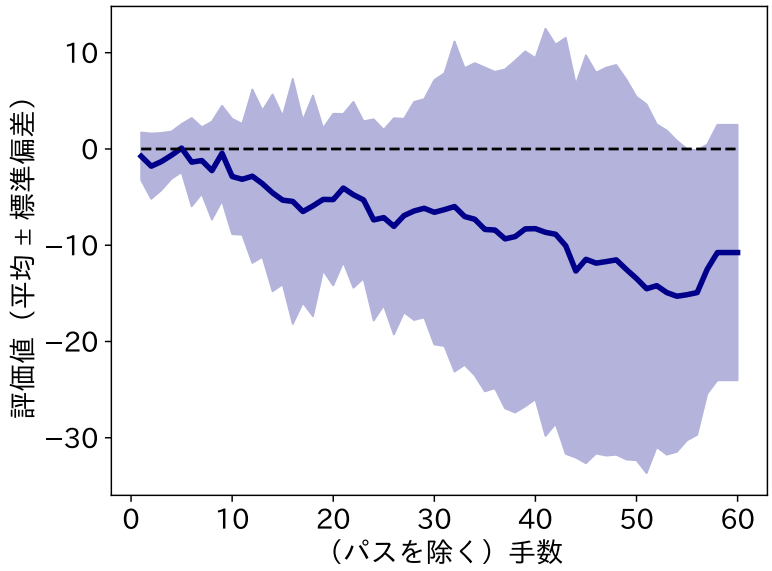

図 5.24　標準偏差 1.0 の乱数を加えた楽観合議プレイヤー vs AlphaZero の局面評価

5.2　楽観合議の実験結果

楽観合議では，勝率の観点からは強さの調整が適切に行われておらず，相手に比べて弱いプレイヤーになっていた．局面評価でも，MCTS1 と 2 では 20 手目以降でマイナスに下がっていく様子が見られた．これは，AlphaZero の評価値の精度が悪く不利な局面で強い手を適切に選択できていないことが考えられる．また，加えた乱数の標準偏差による局面評価の大きな違いは見られなかった．

第 6 章

関連研究

本論文に関連する研究について説明する．

オセロにおける棋力調整の研究として，高木らの研究 [6] では，強さの異なる複数 AI プレイヤーによる棋力調整手法が提案されている．この研究では，異なる複数の AI プレイヤーの打ち手比較によりプレイヤーの強さを対局中に評価し，AI プレイヤーを切り替えることで，ゲームの強さ調整における AI プレイヤーの不自然さの問題解決に取り組んでいる．実験ではオセロを使用し，勝率の拮抗及び手加減による不自然さについて調査している．

第 7 章

まとめ

本研究では，複数の思考プログラムの候補手から一つの手を選択させる合議と呼ばれる手法を用い，より互角に近い手を選択させることで AlphaDDA による動的強さ調整の改良を目指した．

多数決合議と楽観合議による合議プレイヤーの対局結果では，単一の DDA プレイヤーと比較すると極端な評価値の振れは減少したが勝率の面では課題が残った．また，楽観に関しては AlphaZero の評価値に問題がある可能性がある．

これらの問題を解決するために，適切なプレイヤー数や合議に使用する DDA プレイヤーの選別，AlphaZero の長期間の学習などが必要であると考える．

謝辞

本研究を行うにあたって，プログラム作成や論文において数多くの助言を頂いた松崎先生には大変お世話になりました，心より感謝申し上げます．また，高田先生及び竹内先生には副査を引き受けていただき心より感謝申し上げます．

参考文献

[1] K. Fujita, “AlphaDDA: strategies for adjusting the playing strength of a fully trained AlphaZero system to a suitable human training partner.” PeerJ Computer Science 8:e1123, 2022.

[2] 久保田 留奈, 松崎 公紀, “AlphaDDA の局面評価値を用いた再評価”, 電気・電子・情報関係学会四国支部連合大会論文集, 2023.

[3] 伊藤 毅志, 小幡 拓弥, 杉山 卓弥, 保木邦仁, “将棋における合議アルゴリズム-多数決による手の選択”, 情報処理学会論文誌, Vol.52, No.11, pp.3030-3037, 2011.

[4] N. Sephton, P. I. Cowling, N. H. Slaven, “An Experimental Study of Action Selection Mechanisms to Create an Entertaining Opponent.”IEEE Conference on Computational Intelligence and Games (CIG), pp.122–129, 2015.

[5] Edax, `https://github.com/abulmo/edax-reversi`.

[6] 高木 騰也, 藤井 叙人, 片寄 晴弘, “強さの異なる複数の AI エージェントによる オセロのための自然な棋力調整手法の提案”, 情報処理学会論文誌, 63(11),p.1602-1607, 2022.

付録 A

AlphaZero のモンテカルロ木探索

AlphaZero で使用されるモンテカルロ木探索は，木探索アルゴリズムの一つで，ゲームの可能な手の木を効率的に探索し，最適な手を見つけるために使用される．

ゲーム木の各ノードはゲームの状態を表し，全ての合法手に対して辺 (s, a) を持つ．各辺には訪問回数 $N(s,a)$，累計価値 $W(s,a)$，$Q(s,a) = W(s,a)/N(s,a)$，選択確率 $P(s,a)$ が記録されている．

モンテカルロ木探索は「選択」「評価」「展開」「更新」の 4 ステップから構成される．

「選択」において，探索はルートノードから始まり，子ノードが存在したら選択して移動するという操作をリーフノードに到達するまで繰り返す．この時，AlphaZero では次の値が最も大きな子ノードを選択する．

$$a_t = \arg\max_a (Q(S_t,a) + C_{puct} * P(S_t,a) * \frac{\sqrt{N(S_t)}}{(1+N(S_t,a))})$$

$N(S_t)$ は親ノードの訪問回数，C_{puct} は「勝率」と「手の確率 * バイアス」のバランスを調整するための定数である．

AlphaZero の「評価」では，ニューラルネットワークがリーフノードを評価し手の確率と価値を取得する．

その後,「展開」が行われノードの訪問回数が 1 回以上なら子ノードを作成する．ニューラルネットワークを使用することで，複数回シミュレーションを行わなくても有効な手をある程度推論できる．

「更新」では，評価で取得した価値をもとにノードの情報（累計価値と訪問回数）を更新しながらルートノードまで戻る．