置き碁における報酬関数の改良

伊藤　秀将†,柴原　一友††,小谷　善行†

東京農工大学†　テンソル・コンサルティング株式会社††

Abstract:囲碁において多大な効果をあげたモンテカルロ法だが，勝率が極端に偏っていると性能が落ちるという弱点がある．ハンディキャップ前提のゲームである置き碁では，その勝率は初期局面から大きく偏っており，そのため通常の階段型報酬関数では効果を上げにくい．それらを解決する方法として，シグモイド関数による勝敗とスコアの重ねあわせを行う手法と勝率を底上げする Dynamically Adjusting Komi Value という手法が存在する．それらの報酬関数を組み合わせ，より置き碁問題に適した形に改変した．実験の結果，提案手法は従来手法を上回る性能を示した．

Improved Reward Function in Handicap Go

Hidemasa Ito†, Kazutomo Shibahara††,Yoshiyuki Kotani†

Tokyo University of Agriculture and Technology†
Tensor Consulting Co. Ltd.††

Abstract: Monte Carlo method gave a significant effect in the game of go. However, there is weakness such that its performance falls if the winning rate biased extreming. Winning rate in Handicap Go is greatly biased even at the initial phase.Winning percentage of which is large deviation from the initial phase. Therefore, the Monte Carlo method using a staircase reward function　is difficult to achieve an effect. In order to solve this problem, there are the two methods: "Combining Final Score with Winning Percentage by Sigmoid Function in Monte-Carlo Simulation" and "Dynamically Adjusting Komi Value". We combined and modified the reward functions into a form which is more suitable for the game go. The experimental results showed the proposed method outperforms them.

1.　はじめに

人工知能の一分野であるゲーム研究において，零和有限確定完全情報ゲームは題材として頻繁に取り上げられる．囲碁もその一つに含まれる．近年，着手の評価にモンテカルロ法を用いる手法が登場し，コンピュータ碁は大きく発展した．[1] [2]モンテカルロ法には局面全体の勝率が極端に偏り各着手の勝率に差がなくなると，有効な手を判断することが出来なくなる弱点が存在する．そのため，置き碁のような勝率の偏りが前提のゲームには非常に弱い．[3]それらを解決する方法として，シグモイド関数によって終局時の目差と勝敗を重ね合わせる手法[4]とコミを与えて勝率を底上げする手法[5]が存在する.本論文は，それらの報酬関数を組み合わせることで，置き碁におけるモンテカルロ法の性能の向上を試みた．

2.　関連研究

通常のモンテカルロ碁では，プレイアウトの結果，勝利に 1 敗北に 0 の報酬を与えるが，それ以外に，スコア情報などを報酬に含める手法も存在する．本章では勝率の偏った局面に特に有効であると思われる二つの手法を紹介する．

2.1　シグモイド関数によるスコアと勝敗の重ね合わせ

モンテカルロ法の特徴であった，僅差で勝ち，大差で負ける性質を改善することを目的に，柴原らによって提案された手法である［4］．それまでの報酬関数はプレイアウト終局時の目差に対して 0 か 1 の二値を返す階段関数型になっていたが，以下の式のように報酬関数にシグモイド関数を使った．

$$f(x) = \frac{1}{1 + \exp^{-kx}}$$

x:プレイアウトのスコア
k:ゲイン

勝敗価値に終局時の目差を重ね合わせ，よりスコアの高い局面を目指してうつことが出来る．この手法に

より，勝率が高いか低いするとき，勝負を投げたような悪手を打ってしまう現象がある程度改善された．

2.2 Dynamically Adjusting Komi Value (DAKV)

ゲーム分野においてモンテカルロ法が最も効果を発揮するのは，勝率が 50%程度の局面である．報酬の平均値が中央に近いほど，勝率の高い手と低い手の報酬に大きな差がうまれ，正確に判別できるからである．Petr Baudis らが提案した Dynamically Adjusting Komi Value（以後 DAKV とする）は，勝率が極端に偏っている際，プレイアウトにコミを与えることによって，勝敗の基準をずらし，報酬の平均値が比較的 0.5 に近づくように調整する手法である[5].

$$f(x) = 1 \quad (x \geq c)$$
$$f(x) = 0 \quad (x < c)$$

x:プレイアウトのスコア
c:コミ

3. 提案手法

3.1 事前実験

事前に，シグモイド関数と階段関数を組み合わせた以下の式で表される評価関数を作成した．

劣勢時：

$$f(x) = \frac{1}{1+\exp^{-kx}} \quad (x < 0)$$
$$f(x) = 1 \quad (x \geq 0)$$

優勢時：

$$f(x) = \frac{1}{1+\exp^{-kx}} \quad (x > 0)$$
$$f(x) = 0 \quad (x \leq 0)$$

x:プレイアウトのスコア
k:ゲイン

作成した関数について，以下の条件で実験を行った．

1：九路盤を用いる
2：黒側に 13 子の置き石を置く
3：コミは半目
4：白側の持ち時間は 1.5 秒
5：互いに打つことが出来なくなったとき終局
6：いずれの組み合わせも 500 局対戦
7：提案手法及び従来手法は白側

黒側のプレイヤーは以下の 3 つである．
①ランダム着手
②持ち時間 0.025 秒の原始 MC 法
③持ち時間 0.05 秒の原始 MC 法

劣勢時の実験結果を表 1 に示している．提案手法が黒側であったときも同様の実験を行い，勝率の向上を確認した．

3.2 提案手法

本章では，DAKV のような階段関数を移動させた関数とシグモイド関数を組みわせ，より置き碁に適した報酬関数の形を考案する．今回実験で用いた環境（9 路盤に 13 子の置き石）において，白側初手で 1 万回プレイアウトした際のスコアの分布を，図 1 に棒グラフで示した．これらのプレイアウトに対し，DAKV を用いて，34 目のコミを与えたとする．すると，34 目半差以上の敗北は報酬 0，33 目半差以下の敗北には報酬 1 を与えるようになる．しかし実際には，両者の間には半目差勝利と半目差敗北のようなルールに則った明確な差異はない．よって，本来価値の隔たりは殆どないはずである．また逆に，33 目半差の敗北と半目差の敗北には共に 1 の報酬が与えられるが，明らかに価値に違いがあるはずである．

表 1　半階段シグモイド評価関数（白）の対局結果

	ランダム着手	0.025 秒 MC	0.05 秒 MC
階段関数	23.6(±3.7)%	4.4(±1.8)%	0.0%
ゲイン 0.05 シグモイド	66.6(±4.1)%	8.2(±2.4)%	2.4(±1.3)%
ゲイン 0.1 シグモイド	**73.4(±3.9)%**	14.8(±3.1)%	6.0(±2.1)%
ゲイン 0.2 シグモイド	61.0(±4.3)%	14.8 (±3.1)%	5.8(±2.0)%
ゲイン 0.4 シグモイド	55.8(±4.4)%	15.2(±3.1)%	6.2(±2.1)%
ゲイン 0.8 シグモイド	50.6(±4.4)%	14.8 (±3.1)%	**8.4(±2.4)%**
ゲイン 1.6 シグモイド	51.2(±4.4)%	**16.2(±3.2)%**	7.6(±2.3)%
ゲイン 3.2 シグモイド	49.4(±4.4)%	13.0(±2.9)%	6.4(±2.1)%
ゲイン 0.05 半階段シグモイド	74.8(±3.8)%	17.0(±3.3)%	8.0(±2.4)%
ゲイン 0.1 半階段シグモイド	**75.8(±3.8)%**	**22.2(±3.6)%**	**13.6(±3.0)%**
ゲイン 0.2 半階段シグモイド	64.4(±4.2)%	17.6(±3.3)%	9.0(±2.5)%
ゲイン 0.4 半階段シグモイド	54.2(±4.4)%	16.0(±3.2)%	8.4(±2.4)%

そこで，本論文では階段状になっていた DAKV の報酬関数をシグモイド関数の形に変更したものを提案する．また，事前実験により劣勢時は勝利報酬を 1，優勢時は敗北報酬を 0 に固定したほうが強力であると解ったため，それも同時に適用した．作成した評価関数は次の式で表される．

劣勢時：

$$f(x) = \frac{1}{1 + \exp^{-k(x-c)}} \quad (x < 0)$$
$$f(x) = 1 \quad (x \geq 0)$$

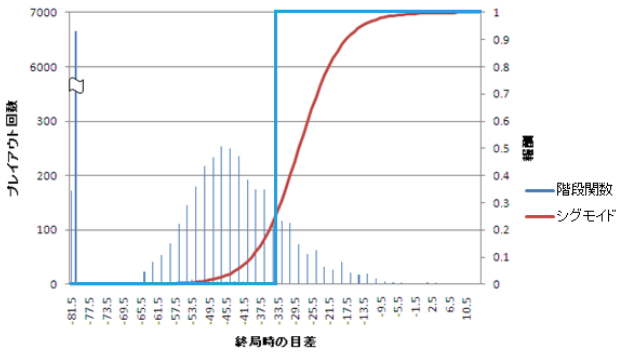

図 1 階段関数とシグモイド関数によるプレイアウト価値

優勢時：

$$f(x) = \frac{1}{1 + \exp^{-k(x-c)}} \quad (x > 0)$$
$$f(x) = 0 \quad (x \leq 0)$$

x:プレイアウトのスコア
k:ゲイン
c:コミ

c の値は，500 回プレイアウトした時の報酬の割合が一定値になるように，一手ごとに決定する．

コミを 34 目とした際の，DAKV と提案手法が与える報酬量を，図 1 に線グラフで示した．

4. 実験

提案手法の有効性を調べるため，劣勢時と優勢時に分けて置き碁の対局実験を行った．

4.1 劣勢時用評価関数の実験

まず劣勢時のものを実験した．コミの値は同条件の事前実験で最も性能が高かった，一手ごとに報酬の平均を 0.05 以上に調整するものを採用した．その他実験の条件は 3.1 の事前実験のとおりである．

実験結果を表 2 に示す．思考時間 0.05 秒のモンテカルロプレイヤーに対する対局では，従来手法の勝率には届かなかったが，思考時間 0.025 秒のモンテカルロプレイヤーに対しては勝率を 37.6(±4.2)%から 47.6(±4.3)%へ，ランダムプレイヤーに対しては 88.6(±2.8)%から 93.6(±2.1)%へ伸ばすことに成功した．

表 2　9 路盤に 13 子の置き石をした場合の，三段階の強さの黒に対する提案手法と従来手法（白）の勝率

	ランダム着手	0.025 秒 MC	0.05 秒 MC
階段関数	23.6(±3.7)%	4.4(±1.8)%	0.0%
ゲイン 0.05 シグモイド	66.6(±4.1)%	8.2(±2.4)%	2.4(±1.3)%
ゲイン 0.1 シグモイド	73.4(±3.9)%	14.8(±3.1)%	6.0(±2.1)%
ゲイン 0.2 シグモイド	61.0(±4.3)%	14.8 (±3.1)%	5.8(±2.0)%
ゲイン 0.4 シグモイド	55.8(±4.4)%	15.2(±3.1)%	6.2(±2.1)%
ゲイン 0.8 シグモイド	50.6(±4.4)%	14.8 (±3.1)%	8.4(±2.4)%
ゲイン 1.6 シグモイド	51.2(±4.4)%	16.2(±3.2)%	7.6(±2.3)%
ゲイン 3.2 シグモイド	49.4(±4.4)%	13.0(±2.9)%	6.4(±2.1)%
階段関数＋コミ	88.6(±2.8)%	37.6(±4.2)%	**24.0(±3.7)%**
ゲイン 0.2 シグモイド＋コミ	91.2(±2.5)%	27.8(±4.0)%	12.8(±2.9)%
ゲイン 0.4 シグモイド＋コミ	**93.6(±2.1)%**	32.4(±4.1)%	16.4(±3.2)%
ゲイン 0.8 シグモイド＋コミ	91.8(±2.4)%	42.2(±4.3)%	17.6(±3.3)%
ゲイン 1.6 シグモイド＋コミ	91.2(±2.5)%	**47.6(±4.4)%**	18.8(±3.4)%
ゲイン 3.2 シグモイド＋コミ	91.0(±2.5)%	34.4(±4.2)%	18.6(±3.4)%

表 3　0.025 秒 MC 法の UCT(階段関数＋報酬コミ)への勝率

	UCT(階段関数＋報酬コミ)
階段関数＋報酬コミ	56.0(±4.4)%
0.05 シグモイド＋報酬コミ	84.8(±3.1)%
0.1 シグモイド＋報酬コミ	**85.6(±3.1)%**
0.2 シグモイド＋報酬コミ	83.6(±3.2)%
0.4 シグモイド＋報酬コミ	74.4(±4.3)%
0.8 シグモイド＋報酬コミ	65.6(±4.4)%

表 4　0.05 秒 MC 法の UCT(階段関数＋報酬コミ)への勝率

	UCT(階段関数＋報酬コミ)
階段関数＋報酬コミ	78.6(±3.6)%
0.05 シグモイド＋報酬コミ	93.2(±2.2)%
0.1 シグモイド＋報酬コミ	**96.0(±1.7)%**
0.2 シグモイド＋報酬コミ	94.4(±2.0)%
0.4 シグモイド＋報酬コミ	93.2(±2.2)%
0.8 シグモイド＋報酬コミ	91.6(±2.4)%

4.2　優勢時用評価関数の実験

優勢時用の評価関数の有効性を調べるため，作成した評価関数を置き碁の黒側に適用し，対局実験を行った．コミの値は同条件の事前実験で最も性能が高かった，一手ごとに報酬の平均を 0.5 以下に調整するものを採用した．

黒側のプレイヤーは以下のものである．それぞれ持ち時間 0.025 秒と 0.05 秒で実験した．

①原始 MC 法（階段関数＋報酬コミ）
②原始 MC 法（ゲイン 0.05 シグモイド＋報酬コミ）
③原始 MC 法（ゲイン 0.1 シグモイド＋報酬コミ）
④原始 MC 法（ゲイン 0.2 シグモイド＋報酬コミ）
⑤原始 MC 法（ゲイン 0.4 シグモイド＋報酬コミ）
⑥原始 MC 法（ゲイン 0.8 シグモイド＋報酬コミ）

白側のプレイヤーは以下のものである．

・UCT（階段関数+報酬 0.05 コミ）

その他条件は 4.1 の実験と同様である．

実験結果を表 3，及び表 4 に示す．優勢時用の評価関数を黒（持ち時間 0.025 秒の MC 法）に適用し，白を階段関数にコミを与える手法にして実験した．すると，持ち時間 0.025 秒では勝率が 56.0(±4.4) %から 85.6(±3.1)%に，0.05 秒では 78.6(±3.6)%から 96.0(±1.7)%に向上した．

5.　実験結果の考察

実験結果より，どちらの評価関数も置き碁においては有効であると示すことが出来た．

優勢な局面での使用は劣勢な局面での使用より効果があがっている．これは報酬平均が 0.05 を基準としたコミよりも，報酬平均 0.5 となるコミの方が，より多くのプレイアウト結果にシグモイド関数の影響を大きく与えられたからだと考えられる．

表 1 を見ると，相手の性能が向上するにつれ，最も勝率の高いゲインが高まっているのが解る．図 2，図 3 は，勝率 0.05 にコミを調整した際のそれぞれの得る報酬の積みあげ図である．赤は報酬総量の半分を初めて超えた目差である．図 2 と図 3 を比較すると，階段関数では 13.5 目勝ちから 25.5 目負けまでのプレイアウトで報酬の半分を得ているのに対し，シグモイドでは 28.5 目となっている．つまり，シグモイド関数を用いると，報酬総量のうち上位のプレイアウトが占める報酬量が階段関数に比べ減るのである．その差が最適なゲインの差となって現れていると言える．

6.　今後の展望

現在は，今回作成した評価関数を平手に応用する研究を行なっている．一定数のプレイアウトの勝率から，現状に適したコミによる勝率の補正度合いとシグモイド関数のゲインを出力する関数を機械学習することで，囲碁の平手対局での勝率向上を行っている．

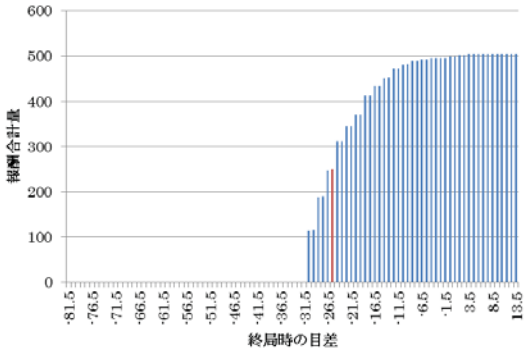

図 2 コミ+階段関数初手の報酬積み上げ図

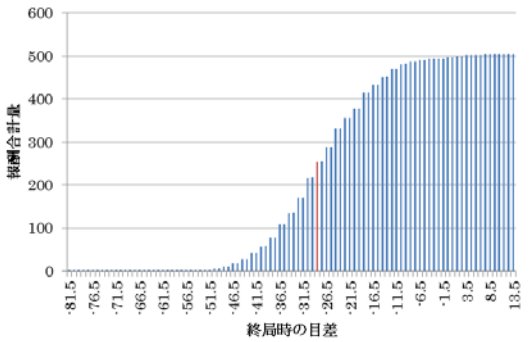

図 3 コミ+シグモイド関数初手の報酬積み上げ図

7.　参考文献

[1] 美添一樹:”コンピュータ碁におけるモンテカルロ法（理論編）”，エンターテイメントと認知科学研究ステーション第5 回招待公演発表資料，2008.

[2] Bernd Br¨uegmann: “Monte Carlo Go，Technical report”， Physics Department，Syracuse University， unpublished， 1993.

[3] 秋山晴彦，是川空，小谷善行:”ゲームにおけるモンテカルロ着手選択の動的勝率調整”，第 52 回プログラミングシンポジウム， pp. 111‐118，2011.

[4] Kazutomo SHIBAHARA， Yoshiyuki KOTANI: “Combining Final Score with Winning Percentage by Sigmoid Function in Monte-Carlo Simulations”，IEEE Symposium on Computational Intelligence and Games(4th CIG2008)， pp. 183‐‐190， Dec. 2008.

[5] Petr Baudis:"Balancing MCTS by Dynamically Adjusting Komi Value，" http://pasky．or．cz/~pasky/go/， 2010.