2F5-5

# ゲーム木探索における満足化価値関数の有効性

## Effectiveness of a Satisficing Value function in Game Tree Search

大用 庫智[*1]
Kwansei Gakuin University

Monte Carlo tree search methods (especially the UCT algorithm), which are proven effective for game AIs, have come into action for real-time games. In that type of games, it is necessary to quickly find appropriate options under extremely limited thinking time. For this reason, the use of UCT is not always appropriate because it requires many simulations to perform well. This study investigated a *satisficing* tree search algorithm. Implementing satisficing behavior with only a simple value function, I show the algorithm behaves according to satisficing. By tuning a single intuitive parameter, the satisficing search enables fast search of the optimal action in a huge search space.

## 1. はじめに

2006 年に囲碁ゲーム AI 研究を劇的に進展させたことで，モンテカルロ木探索（MCTS）に注目が集まった．それ以降，MCTS は囲碁以外のゲーム（e.g., さめがめ，ハーツ，Lines of Action）やその他の問題（e.g., 制約充足問題，スケジューリング）に応用され，その汎用性の高さが示されている[Browne 12]．MCTS はリアルタイムゲーム（e.g., Ms. Pac-Man）への応用も初められている．また近年では膨大な対局データを Deep leaning の技術により学習した MCTS を用いた AlphaGo[Silver 16]が，人間の囲碁の世界チャンピオンに勝利した．MCTS は非常に広い研究領域で扱われている．

MCTS はランダムなサンプリングから計算する勝率を利用することで各ゲームに容易に実装可能である．多くの実装例では，その勝率の代わりに UCB (upper confidence bound) という価値を行動の価値として用いる UCT (UCB applied to trees)が主に利用されてきた．UCT は長期的な運用が可能であればいつかは最適解に到達するが，初期の振る舞いによる指標のぶれと膨大な試行回数が問題になっていた．近年，その代替案として人間の認知の偏りを実装した緩い対称性（LS）モデル[大用 15a]という価値関数を行動の価値として用いる LST (LS model applied to trees)が提案されている[Oyo 15]．LST は，人間認知に観られる「受容可能な基準を満たす選択肢の探索」という満足化[Simon 56]を効率的に行う MCTS の一種であり，後に述べる人工的なゲーム木において適切な満足化基準下であれば UCT よりも高成績を示した．

MCTS は非常に限定された思考時間の中で次の着手を決めなければならないリアルタイムゲームへの応用が進んでおり，より計算量が少ない方法が望まれている．例えば Ms. PacMan では約 60ms に 1 回の行動選択の中で，如何に得点を稼げるかが競われている．高成績を目指すゲーム AI 開発・研究が進む一方で，例えば，約 40ms に 1 回の行動選択を行わなければならない Mario AI では，人間らしくプレイする AI 構築に MCTS が用いられている[中野 13]．不完全な情報下での AI の開発でも高成績以外の方向でも研究が進められている．

そこで本研究では，満足化の単純な価値関数である RS (reference satisficing) モデル[高橋 15]を木探索に応用した RST (RS model applied to trees)[大用 15b]の有効性を検証する．そして，RST が簡単な価値関数を行動価値としてするだけで，木探索においても満足化基準の変化により満足化の振る舞いが現れ，そして，広い探索空間において適切な満足化基準のもとで高成績に繋がることを示す．

連絡先：大用庫智，関西学院大学 総合政策学部，〒669-1337 兵庫県三田市学園 2 丁目 1 番地，kuratomo.oyo[at]gmail.com

## 2. 人工的なゲーム木

Kocsis と Szepesvári は人工的なゲーム木を利用して，UCT が最善手に収束する事を理論とシミュレーションの観点から示した[Kocsis 06]．本研究でも[Kocsis 06, Oyo 15, 大用 15b]と同様に図 1 のようなゲーム木を用いる．このゲーム木では，MAX と MIN の二人のプレイヤーが交互にゲームを進める．MAX と MIN は自身にとって最も有望な着手を選択する．ゲーム終局を意味する葉ノードには，ゲーム終局の評価値が一定の範囲から一定の確率で与えられる．この評価値が一定の基準よりも高ければ MAX の勝利となり，その基準以下であれば MAX の敗北となる．

二人のプレイヤーの着手の行動価値は葉ノードに到達する事で得られる勝ち（1）負け（0）の情報から計算される．そして，現在の局面（ノード）からその直接の子ノードの行動価値が高い方を選択し，葉ノードまで交互に着手を繰り返す．この行動により勝ち負けの情報をサンプリングする．本来はランダムにゲーム終局まで着手を繰り返すことをプレイアウトと呼ぶが，本研究では上記のサンプリングをプレイアウトと呼ぶ．

### 2.1 UCT

プレイアウトと実際の着手選択を行うために，UCT ではサンプリングを工夫するアルゴリズム UCB1 [Auer 2002] が利用される．このアルゴリズムはほぼ価値関数にすぎないが，十分な選択回数が許されれば高い成績を示し，期待損失の上界を保証（即ち最適解収束の保証）している．UCB1 アルゴリズムは最初

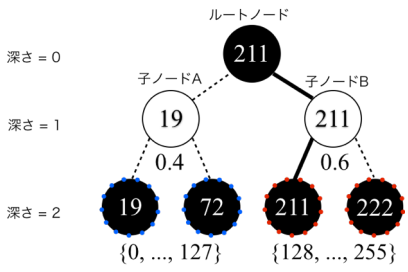


図 1：深さが 2，一つの親ノードが持つ子ノードの数が 2 のゲーム木．赤と青の点線はそれぞれ MAX プレイヤーの勝ち負けを意味している．$P_A$ と $P_B$ はそれぞれ 0.4 と 0.6 である．実線は Mini-Max 的に最適な選択経路である．

に全ての子ノードを選択しなければならない．その後，価値関数UCB$(i,j)$の値が最も高い子ノードを選択する．

$$\mathrm{UCB}(i,j) = \bar{X}_{i,j} + \sqrt{2\ln n_i / n_{i,j}}, \qquad (1).$$

ここで，$\bar{X}_{i,j}$は深さ $i$ の子ノード $j$ の期待値（条件付き確率 $P(1|j)$ と一致）であり，$n_{i,j}$は深さ $i$ の子ノード $j$ の選択回数，$n_i$ は（$j$に限らず）深さ $i$に到達した選択回数を意味する．

## 2.2 RST

価値関数 UCB の代わりに，満足化価値関数である RS モデル [高橋 15] を用いるモンテカルロ木探索を提案した[大用 15b]．RST は以下の RS の値が最も高い子ノードを選択する．

$$\mathrm{RS}(1|\mathrm{A}) = (a+d)/(a+b+c+d) \qquad (2).$$

ここで，$A$と$B$は図 1 の様な子ノード A と B に対応する．共起情報$a$は1と$A$が共に発生した頻度を意味する$N(A,1)$である．同様に$b,c,d$はそれぞれ$N(A,0)$と$N(B,1)$, $N(B,0)e$を意味する．満足化基準 R を用いた RS は$(2\mathrm{R}a + 2\mathrm{R}d)\ /\ (2\bar{\mathrm{R}}(a+c) + 2\mathrm{R}(b+d))$である．R, $\bar{\mathrm{R}}$ は $\mathrm{R} \in [0,1]$, $\mathrm{R} = 1 - \bar{\mathrm{R}}$を満たす．

本研究では式(2)の RS の行動価値を以下のように書き換えた式(3)を MCTS の行動価値関数として利用する．

$$\mathrm{RS}(i,j) = n_{i,j}(\bar{X}_{i,j} - \mathrm{R}) \qquad (3).$$

# 3. シミュレーション

ここでは[Kocsis 06, Oyo 15, 大用 15b]と同様な設定を用いて，最適解収束の速さとその振る舞いの仕方の観点から RST と UCT を比較する．人工的なゲーム木の深さ (Deep:D) と一つの親ノードが持つ子ノードの数 (Balancing factor:BF) は，3.1 節と 3.2 節でそれぞれ設定される．葉ノードの親ノード X には確率$P_X$が設定されている．葉ノードの評価値が割り振られる範囲は確率$P_X$ で{128, ..., 255}，$1-P_X$ で{0, ..., 127} のどちらかが与えられる．葉ノードの評価値はその範囲から一様に与えられる．このゲーム終局（葉ノード）の評価値が 128 以上であれば MAX が勝利する．即ち，$P_X$が高ければ MAX にとって有利な局面を意味する．この設定に従い 100 種類の木を生成した．そして，それぞれの木に対して 1000 プレイアウトを 100 回実行し，その平均を結果とした．指標は正解率と切り替え率の二つを用いる．正解率は Mini-max 的に最適なノードを選択した割合である（具体例は図 1 参照）．切り替え率は前回の選択から選択肢を変えた割合である．UCT には UCB の価値に∞を代入して初期方策を実現する．

[大用 15b]では，R が 0.5 の式(2)を用いて，高確率環境 ($P_A$=0.8, $P_B$ = 0.6) と単高確率環境 ($P_A$=0.6, $P_B$ = 0.4)，低確率環境 ($P_A$=0.4, $P_B$ = 0.2) の三種類の確率設定毎に D=20, BF=2 の結果を示し，RST の振る舞いが基準 R=0.5 に深く関連していることを示した．つまり，RST は満足化基準よりも価値が高いと判断する腕がある場合 (高確率と単高確率) には探索をほぼ停止し，それよりも価値が低いと判断する選択肢 (低確率)しか無い場合は探索を多く行った．単高確率において RST は UCT よりも高い性能を示した．式(3)を用いた RST でも[大用 15b]と同様な結果を得られた．

## 3.1 満足化基準の可変化

そこで RST の R の変化に伴う振る舞いを観測するために，確率の設定（$P_A$, $P_B$）は(0.4, 0.2) と(0.6, 0.8) とした．(0.4, 0.2) の場合の R は 0.5, 0.3, 0.1, (0.6, 0.8) の場合の R は 0.9, 0.7, 0.5 とした．RST の満足化基準を変化させた．D と BF はそれぞれ 20 と 2 とした．図 2 の全ての確率が R 未満の環境（$P_A$=0.8, $P_B$=0.6, R = 0.9 と$P_A$=0.4, $P_B$=0.2, R = 0.5）では，RST は切り替え率が高い．図 2 の確率が一つでも R 以上の環境では，RST は切り替え率が低い．図 2 の二つの確率の間に R が設定されている環境では，RST は UCB1 よりも成績が高い．

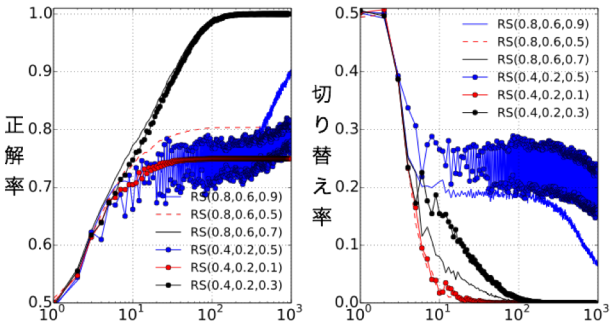


図 2：基準を変化させた RS の正解率（左）と切り替え率（右）．

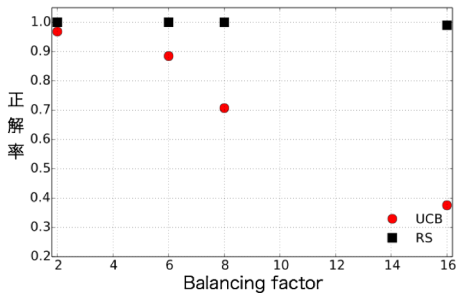


図 3：一つの親ノードが持つ子ノードの数毎の UCB と適切な満足化基準下の RS の正解率．

## 3.2 広い探索空間における満足化

RST のスケーラビリティ性を示すために，探索空間が異なる人工的な木を 4 種類用いる．BF と D は，それぞれ (2, 24) と (4, 12), (8, 8), (16, 6) とした．全てのノードの数は $\{2^{(24+1)}-1, 4^{(12+1)}-1, 8^{(8+1)}-1, 16^{(6+1)}-1\}$ である．$P_1, P_2, ..., P_N$ は 0.6, 0.4, ..., 0.4 である．この設定では，適切な満足化基準として，$R$ は 0.5 とした．図 3 には 1000 回後のプレイアウトの正解率を示す．図 3 から RST の成績は UCT よりも成績が高いことが分かる．

# 4. 議論と結語

本研究では満足化価値関数 RS を用いた MCTS が満足化基準に従い，その振る舞いに変化が見られた．基準 $\mathrm{R} < P_X$ となる選択肢が一つでもあれば，満足する選択肢を発見したため，RST は選択肢への切り替えをほぼ停止させる（図 2 右）．また，基準 $\mathrm{R} < P_X$となる選択肢が一つでもなければ，満足する選択肢を見つけるために，RST は選択肢への切り替え率が高いままとなる（図 2 右）．このように一定の基準 R に至る選択肢を見つけると探索をやめるという満足化の行動が MCTS においても同様に見られた．

UCT と RST の探索方法には異なる傾向がある．UCT は「不確実な時には楽観的に」という原理[Bubeck 12]に従い未知のノードを優先的に訪問するため，探索木の横幅を比較的広く探索する傾向がある．一方で RST は満足化に従い基準以上のノードを優先的に訪問するため，探索木を比較的深く探索する傾向ある．木探索を用いる実例では探索方法に優劣があるため，問題設定毎に価値関数を変更することが重要であると考える．

図 2 左から満足化は非最適化行動を行うことがあるが，最善手と次善手の間であれば満足化は最適化と等しくなる．この R の設定は，本来では知りえない情報を用いて静的に設定されて

いる. しかし, 強化学習の一般的な手法では, 局所解から脱出や探索のためのパラメーターを試行錯誤で決めなければならない場合がある. それらと比較すると, 高成績のためには最善手と次善手の間を取れば良く, この意味では満足化は直感的である.

これらの満足化の振る舞いは, ゲーム AI の強さの容易な調節や広大な探索空間から目的の選択肢を見つけることに有効に働くと考えられる.前者の例として, これまでシミュレーションで用いてきた確率と基準を勝率と対応付ければ, 基準 R を低めに設定した RST は, 最適な選択肢を見つける前に基準以上の選択肢で満足するために探索を停止する比較的弱い AI の作成が可能である. 同様に後者の例として, 一定の目的(e.g., 勝率)以上の選択肢を巨大な探索空間から探すことができる(切り替え率の低下が発見の合図). このように満足化は単に価値関数を切り替えるだけで UCT は異なる方法として利用可能である. 今後は, 基準以上の選択肢を発見する割合(満足度)や, 満足化基準の動的な設定方法, 探索方法の定量的な分析などを行い, 満足化の有効性をさらに検証する.

## 参考文献


[Auer 02] Auer, P., Cesa-Bianchi, N., and Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem, *Machine learning*, 47, 23–256 (2002).

[Browne 12] Browne, C., Powley, E. Whitehouse, D., Lucas, S., Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., Colton, S.: A survey of Monte Carlo tree search methods, *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1), 1–43 (2012).

[Bubeck 12] Bubeck, S and Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems, *Foundations and Trends in Machine Learning*, 5(1), 1–122 (2012).

[Oyo 15] Oyo, K., and Takahashi, T., Efficacy of a causal value function in game tree search, *International Journal of Parallel, Emergent and Distributed Systems* (2015).

[大用 15a] 大用庫智, 市野学, 高橋達二:緩い対称性を持つ因果的価値関数の認知的妥当性と N 本腕バンディット問題におけるその有効性, 人工知能学会論文誌, 30(2), 403–416 (2015).

[大用 15b] 大用庫智, 高橋達二:ゲーム木探索における満足化の効果, 2015 年度人工知能学会全国大会(第 29 回)予稿集, 2L1-2 (2015).

[Simon 56] Simon, H. A.: Rational choice and the structure of the environment, *Psychological Review*, 63(2), 129–138 (1956).

[Silver 16] Silver, D et al., Mastering the game of Go with deep neural networks and tree search, *Nature*, 529, 484–489 (2016)

[Kocsis 06] Kocsis, L. and Szepesvári, C.: Bandit based Monte-Carlo Planning, *Machine Learning: ECML 2006 In Proceedings of the 17th European conference on Machine Learning*, 4212, 282–293 (2006).

[中野 13] 中野 雄基, 美添 一樹, 脇田 建: 進化計算と UCT による Mario を人間らしくプレイする AI, ゲームプログラミングワークショップ 2013 論文集,81–88 (2013).

[高橋 15] 高橋達二, 甲野祐, 大用庫智, 横須賀聡:不確実性の下での満足化を通じた最適化, 2015 年度人工知能学会全国大会(第 29 回)予稿集, 2D1-OS-12a-4in (2015).