F-053

状態価値に基づく行動選択によるボードゲーム AI の難易度調整
Difficulty Adjustment in Games via Action Selection Based on Thresholded State Values

坂本 充生 *　　伊原 滉也 *
Mitsuki Sakamoto　　Koya Ihara

1　はじめに

対戦型ボードゲームにおいて AI の強さを適切に制御することは，魅力的なゲーム体験を実現する上で重要な課題である．AI の強さが適切でない場合，挑戦性の欠如やゲームのクリア不能といった問題が生じ，プレイヤーの満足度を損なう要因となる．現在，ゲーム AI の強さは多くの場合開発者によって状態遷移に基づくビヘイビアツリーなどの構造を用いて手動で調整されている [1]．これらの手法は設計時に多くの試行錯誤を必要とする．

一方，強化学習やプランニングに代表される機械学習の手法は，環境における最適な行動方策を獲得することに焦点を当てている [4]．これらの手法は最適な方策に基づく強力な AI を構築可能だが，そのままでは一般プレイヤーにとって難易度が高すぎることが多い．すなわち，最適な行動を実現する AI の構築と，プレイヤーにとって適切な強さを持つ AI の設計とは，目的の異なる独立した問題である．

AI の強さ制御に関する手法は提案されてきたが，多くは経験的調整や特定の手法に依存しており，行動選択と強さとの関係を定量的に制御するための一般的な枠組みは十分に確立されていない．例えば，チェスにおいては探索アルゴリズムの深さを制限することで AI の強さを調整でき，これがプレイヤーのレーティングと相関することが知られている [2]．また，AlphaZero のように汎用的な強化学習アルゴリズムを用いた強さの調整手法も存在するが [3]．これらは依然として特定のアルゴリズム的枠組みに依存している．

本研究では，このような課題を踏まえ，状態価値関数に基づいて行動選択確率を調整することで，AI の強さを連続的かつ安定的に制御可能な行動選択アルゴリズムを提案する．提案手法は，まず目標とする報酬値を指定し，行動選択の際にはその目標値よりも Q 値が大きい行動のうち，最も Q 値の小さいものを選択するという戦略に基づく．Q 値（Q 関数の値）は，ある状態 s において行動 a を選択した場合に将来的に得られる報酬の期待値を表す．もし一度でも目標値より小さい Q 値の行動を選択すると，その後に目標値を達成することができなくなる．したがって，このような方策に従うことで，目標とする報酬を得るための最低限の行動を選択し続けることが可能となる．

本研究では，この手法の有効性を確認するために，単純な三目並べにおいて実験を行い，強さの連続的かつ滑らかな制御が可能であることを示す．

2　従来手法

ボードゲームの多くは，有限の状態集合 S，行動集合 A，遷移関数 T，報酬関数 R を持ち，両プレイヤーが交互に行動するターン制ゲームである．またそのほとんどは，ゲームの状態およびプレイヤーの行動はすべてのプレイヤーに観測可能であるため，これらは 完全情報二人ゼロ和ゲームとして定式化できる．ここでプレイヤー 1 およびプレイヤー 2 の方策をそれぞれ π_1，π_2 とする．

相手プレイヤの戦略 π_{opponent} を固定すると，ゲーム環境は一人称のエージェントが行動する マルコフ決定過程（MDP）として記述可能である．MDP は以下の 5 要素で定義される：$\mathcal{M} = (S, A, P, R, \gamma)$ ここで，$P(s' \mid s, a)$ は状態 s において行動 a を選択したときに遷移する確率であり，$R(s, a)$ は即時報酬，$\gamma \in [0, 1)$ は割引率である．固定された方策のもとで，動的計画法により最適 Q 値関数 $Q^*(s, a)$ を次式で計算できる：$Q^*(s, a) = \sum_{s'} P(s' \mid s, a) \left[R(s, a) + \gamma \max_{a'} Q^*(s', a')\right]$．このようにして，固定方策のもとでの最適行動列が得られ，強化学習アルゴリズムにより学習・計算が可能である [5].

強化学習やプランニングに代表される機械学習の手法は最適行動の導出に焦点を当てており，そのままでは AI の強さを任意に調整することは困難である．このため，AI の強さ制御に関する手法が別途提案されている．チェスなどの古典的ボードゲームにおいては，探索の深さを制限することで AI の強さを調整する手法が用いられている．実際，探索の深さと Elo レーティングの関係は実験的に示されている．[2]．また強化学習に関連する分野では，[3] らは方策をボルツマン方策にして確率化することで強さの調整している．ボルツマン方策では各行動 a の選択確率を，その Q 値に基づいてソフトマックス関数で定義する：$\pi(a \mid s) = \frac{\exp(Q(s,a)/T)}{\sum_{a'} \exp(Q(s,a')/T)}$ ここで，T は温度パラメータであり，$T \to 0$ で最適方策，$T \to \infty$ で一様ランダム方策に近づく．

* サイバーエージェント

Algorithm 1 最小満足行動選択アルゴリズム

Require: 状態 s，Q 関数 $Q(s,a)$，目標報酬 τ

$A_\tau(s) \leftarrow \{a \in A \mid Q(s,a) \geq \tau\}$
if $A_\tau(s) = \emptyset$ **then**
  $a^* \leftarrow \arg\max_{a \in A} Q(s,a)$
else
  $a^* \leftarrow \arg\min_{a \in A_\tau(s)} Q(s,a)$
end if
return a^*

3 提案手法

本研究では，任意の目標報酬 τ を指定し，Q 関数 $Q(s,a)$ に基づき，その目標を上回る行動のうち，可能な限り最小の Q 値を持つ行動を選択する．この手法を本稿では「最小満足行動選択」と呼ぶ．

状態 s において目標報酬 τ を達成可能な行動の集合を以下のように定義する：$A_\tau(s) = \{a \in A \mid Q(s,a) \geq \tau\}$. 次に，この集合に含まれる行動のうち，Q 値が最小となる行動 a^* を選択する：$a^* = \arg\min_{a \in A_\tau(s)} Q(s,a)$. なお，目標報酬を上回る行動が存在しない場合には，最も Q 値の大きい行動を選択する．この選択戦略により，目標値を下回ることなく，その達成に必要最小限の行動を継続的に選択することが可能となる．この手続き全体は，アルゴリズム 1 に示すとおりである．

4 実験

本節では，提案手法である最小満足行動選択アルゴリズムの有効性を検証するために，完全情報ターン制ボードゲームである三目並べを題材として実験を行う．三目並べは 3×3 の盤面上でプレイヤーが交互に手を打つ 2 人ゼロ和ゲームであり，状態空間が有限かつ完全に列挙可能であることから，Q 関数の厳密な算出が可能である．報酬はより少ない手数での勝利を目指すような方策を獲得するために次のようにする．プレイヤーが勝利した場合には $+1.0$ の報酬を得る．加えて，早期の勝利を評価するため，ゲームの各ターンには -0.2 のステップペナルティを課す．敗北プレイヤーには，勝利プレイヤーが得た報酬の符号を反転させた値を与えることで，ゲームの報酬構造をゼロ和に保っている．相手プレイヤーは固定されたルールベースの戦略に従って行動する．具体的には，勝利のチャンスがある場合には必ずその手を選択し，そうでない場合には相手のリーチを阻止することを優先する．この固定方策の相手に対して，動的計画法を用いてすべての状態における最適な Q 関数 $Q^*(s,a)$ を計算し，以後のエージェントにおける行動選択に用いる．比較対象としたエージェントの方策は 2 種類ある．1 つ目は，ボルツマン方策エージェントである．この手法では，Q 関数 $Q^*(s,a)$ に基づき，温度パラメータ T を調整することで行動選択確率をソフトマックス関数で確率的に定義する．2 つ目は，本研究の提案手法（最小満足選択アルゴリズム）であり，目標とする報酬 τ をあらかじめ設定し，それを達成可能な行動の中で最も Q 値が小さい行動を選択する．各設定において 10 エピソードの対戦を行い，得られた累積報酬の平均および標準誤差を算出する．

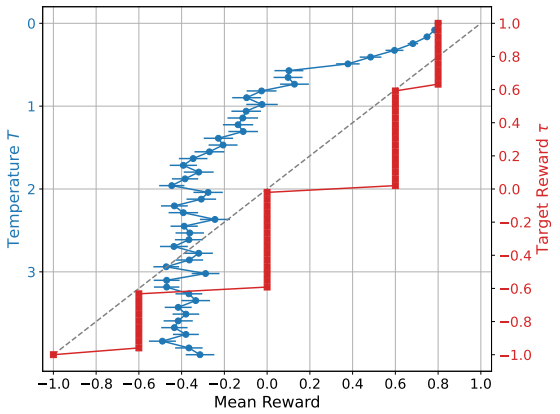

図 1: 温度指定および目標報酬指定による強さ制御手法の平均報酬の比較

図 1 に，両手法における調整パラメータ（ボルツマン方策では温度 T，提案手法では目標報酬 τ）に対する平均報酬の変化を示す．図 1 に示されるように，ボルツマン方策では温度パラメータ T の変化に応じて平均報酬が減衰するものの，その変化は一貫性に欠け，報酬値は非単調である傾向ある．一方，提案手法においては，目標報酬 τ の設定と実際に得られた平均報酬との間に明確な対応関係が確認され，$y = x$ の線上に近い挙動を示している．

この結果は，提案手法が目標とする強さ（累積報酬）を直接かつ定量的に制御可能であることを示しており，確率的なボルツマン方策と比較して，より高い精度での強さ制御が可能であることがわかる．さらに，平均報酬の標準誤差は 0 であり，行動選択の安定性という観点においても，提案手法の優位性が確認できる．

参考文献

[1] M. Colledanchise and P. Ogren. *Behavior Trees in Robotics and AI: An Introduction*. 07 2018.

[2] D. Ferreira and P. Aguiar. The impact of search depth on chess playing strength. In *International Conference on Computer Games*, pp. 1–12, 2013.

[3] K. Fujita. Alphadda: Strategies for adjusting the playing strength of a game artificial intelligence. *PeerJ Computer Science*, 8:e1123, 2022.

[4] D. Silver, T. Hubert, J. Schrittwieser, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[5] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.