第30回世界コンピュータ将棋選手権

Miacisアピール文書

慶應義塾大学 修士2年 迫田真太郎

2020/03/24

Miacisの特徴

- 評価値を確率分布として出力

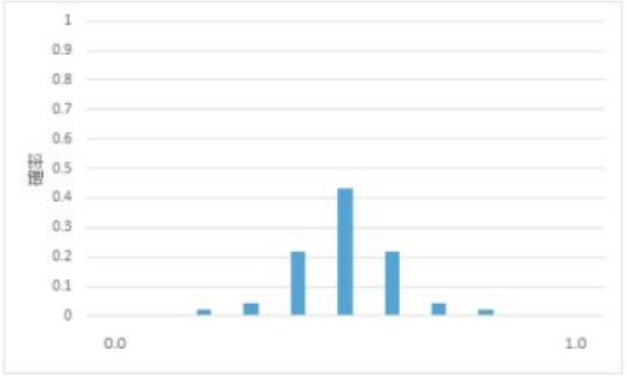
従来の大多数

評価関数

予測勝率 0.55

提案手法

評価関数



背景

- 将棋ソフトにおける評価値≒強化学習における状態価値
- 状態価値は「ある方策πの下での累積報酬の期待値」
- 期待値ではなく分布を直接近似する→分布型強化学習
  - 森村らの研究[1]
    - 分布からリスクを考慮した行動選択
  - Categorical DQN[2]
    - 深層強化学習においてカテゴリカル分布によって累積報酬分布を近似
  - Windfall[3]
    - 形勢の分布を考慮した将棋ソフト

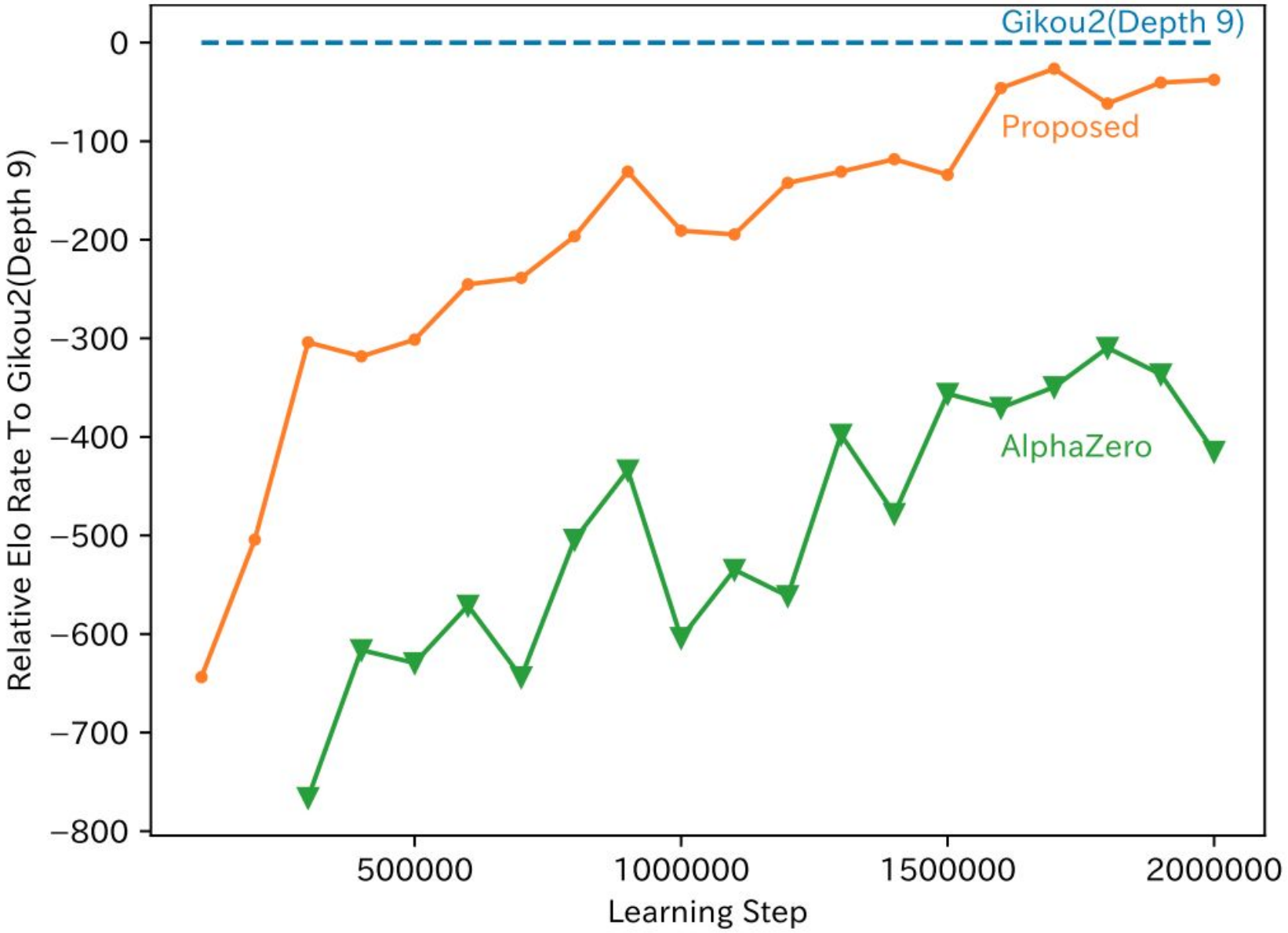
[1] Tetsuro Morimura, et al. “Parametric return density estimation for reinforcement learning.” arXiv preprint arXiv:1203.3497 (2012).
[2] Marc G. Bellemare, Will Dabney, and Remi Munos. “A distributional perspective on reinforcement learning.” Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
[3] 井本康宏, http://denou.jp/tournament2017/img/pr/windfall.pdf

提案手法

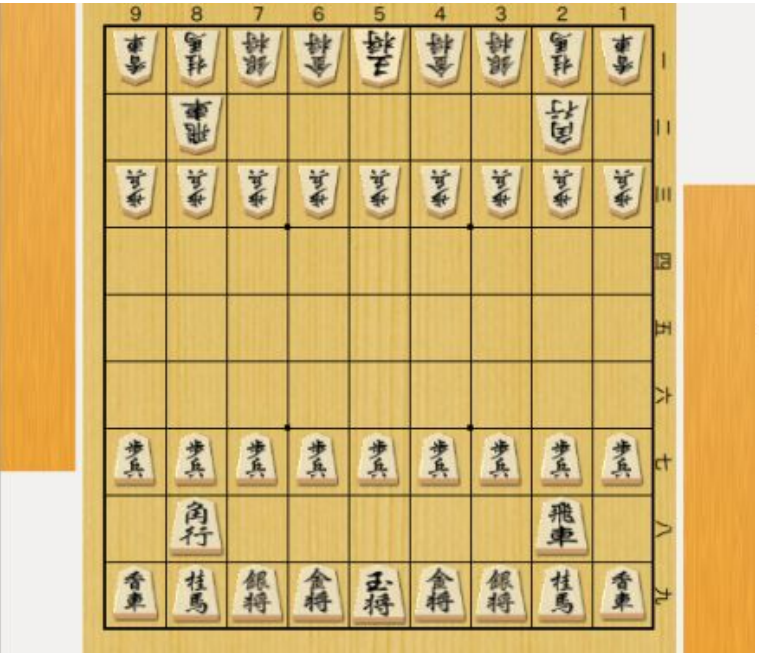
- 評価値を確率分布として出力
  - AlphaZeroモデルのValue部分を複数ユニットに拡張
  - Categorical DQNを参考に報酬[-1, 1]を51分割
    - 状態価値が各区間に含まれる確率を出力
- 確率分布を用いた効率的な探索
  - MCTSの選択ステップで良い価値となる確率を計算

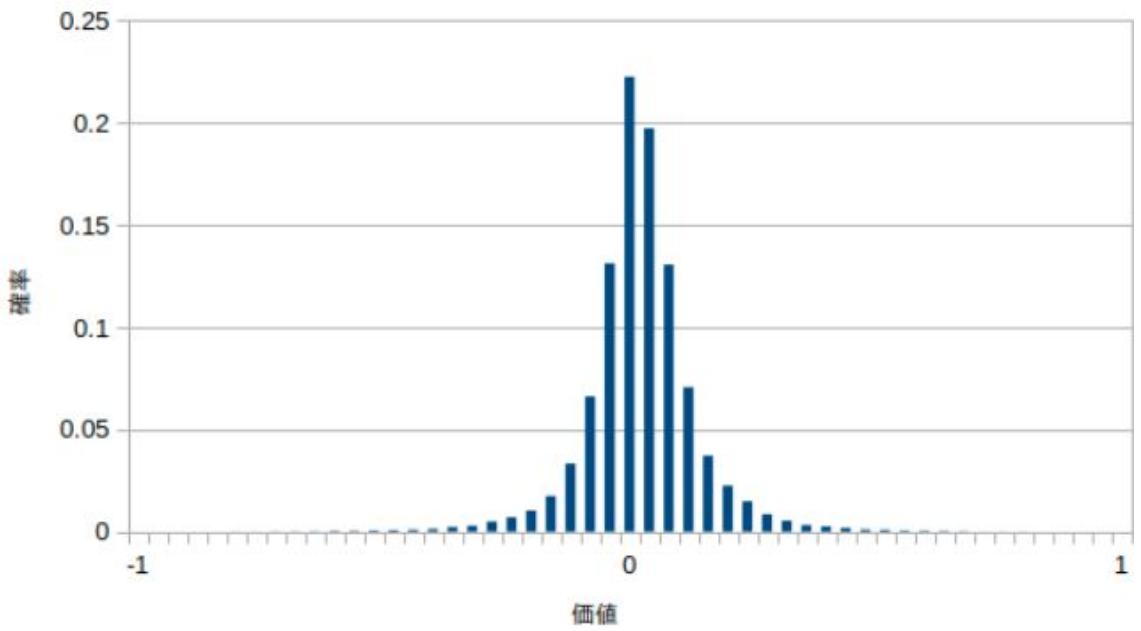
実験結果

- AlphaZeroよりも高い性能



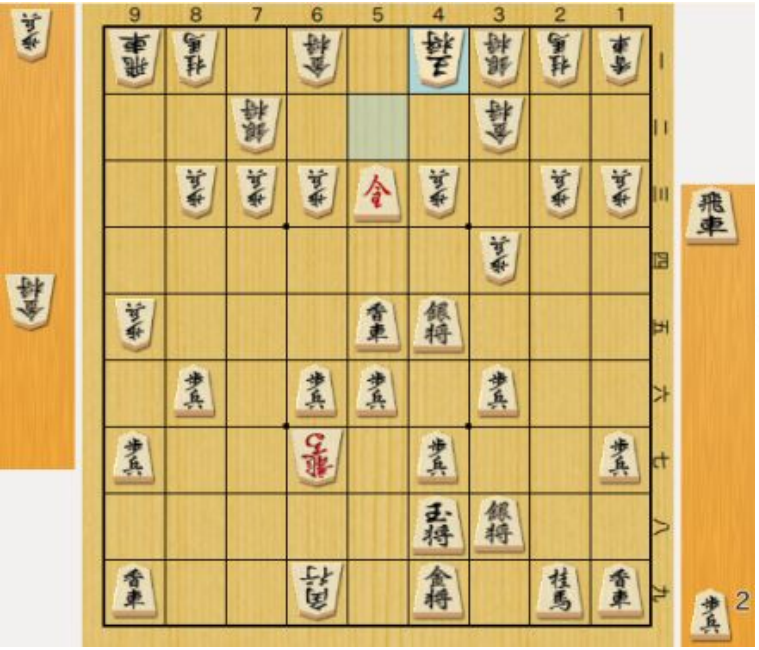
出力例1

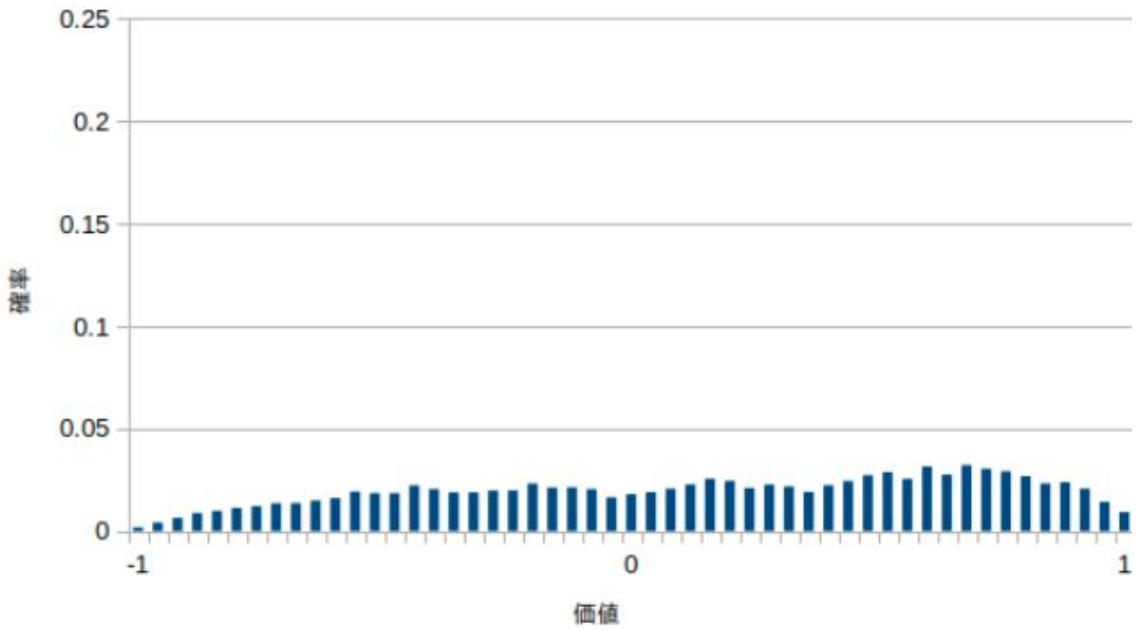




- 初期局面では 0 付近に高い確率を示す

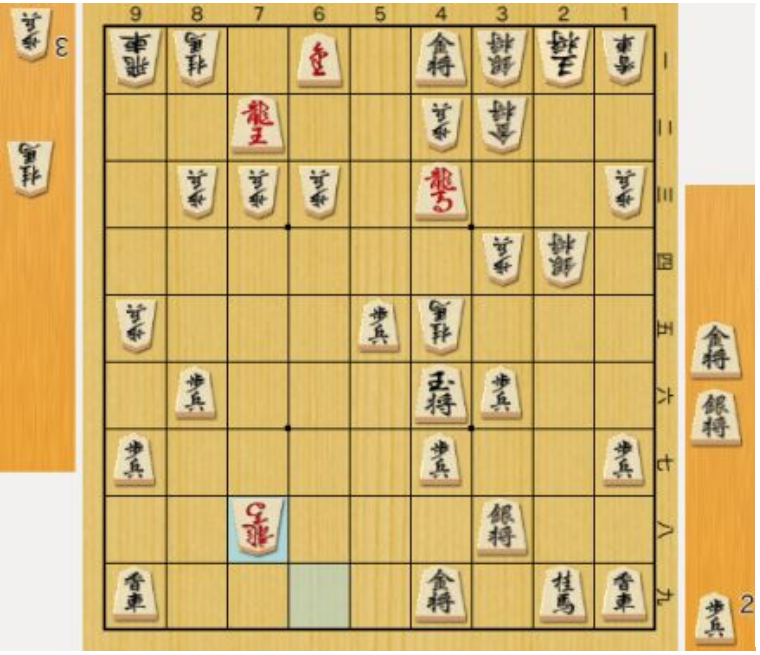
出力例2

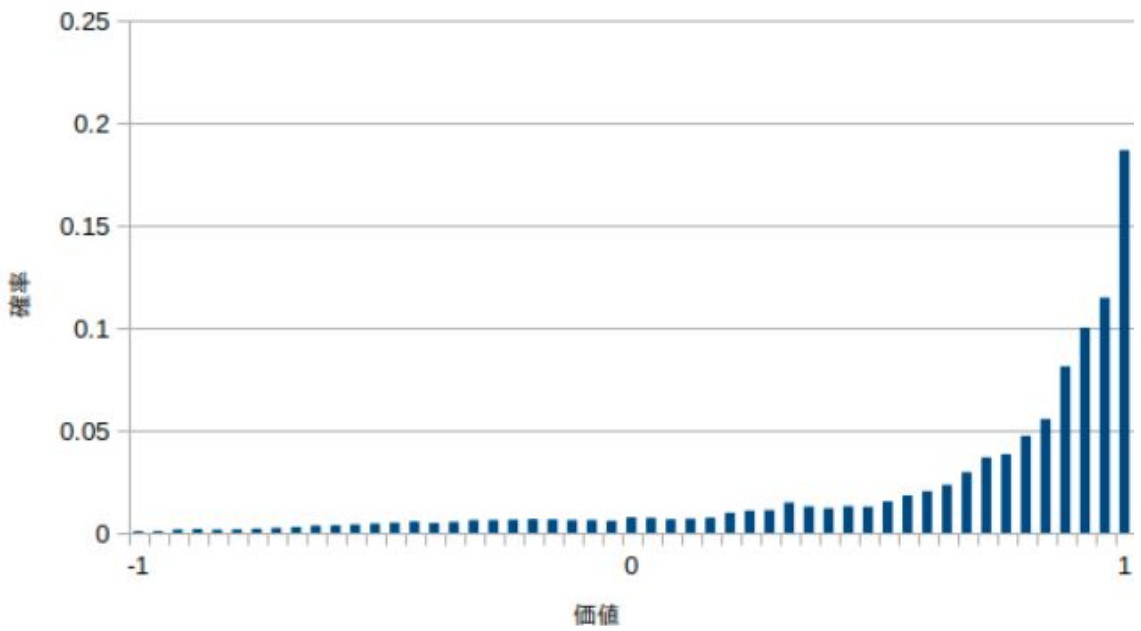




- 中盤の難所 (？) では、期待値はほぼ 0(互角) だが幅広い値に確率を示す
  - 「難解」あるいは「形勢不明」という判断？

出力例3





- 勝ちになった終盤では 1 付近に高い確率を示す