

Miacis アピール文書

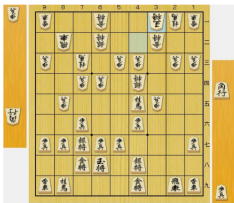
迫田真太郎

April 12, 2019

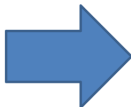
Miacisの特徴

◆ 評価値を確率分布として出力

従来の大多数

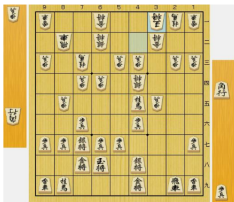


評価関数

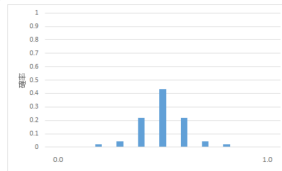
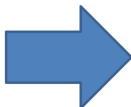


予測勝率 0.55

提案手法



評価関数



背景

- ◆ 将棋ソフトにおける評価値 $\equiv$ 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」
- ◆ 累積報酬の分布を直接近似する $\rightarrow$ 分布型強化学習
 - 森村らの研究ⁱ
 - 累積報酬の分布からリスクを考慮した行動決定を実行
 - Categorical DQNⁱⁱ
 - 深層強化学習においてカテゴリカル分布による累積報酬分布の近似を行う
 - Windfallⁱⁱⁱ
 - 形勢の分布を考慮した将棋ソフト

ⁱTetsuro Morimura, et al. "Parametric return density estimation for reinforcement learning." arXiv preprint arXiv:1203.3497 (2012).

ⁱⁱMarc 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.

ⁱⁱⁱ<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

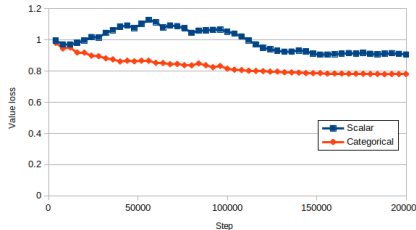
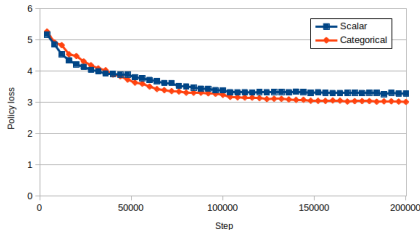
- AlphaZero モデルの Value を複数ユニットに拡張
- Categorical DQN を参考に報酬 $[-1, 1]$ を 51 分割
 - 状態価値が各値となる確率を出力

◆ 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考慮

実験結果 (学習推移)

- ◆ 強化学習により自己対局から分布を学習
- ◆ 分布としたことで評価精度向上
 - floodgate の R2800 以上同士の棋譜に対する損失



実験結果 (対局)

◆ 技巧 2^{iv} (深さ 1, R1506)^v と 1 手 1 秒で 100 局対局

出力	勝率 (%)	推定レート
Scalar(従来手法)	10.0	1124.3
Categorical(提案手法)	72.0	1670.1

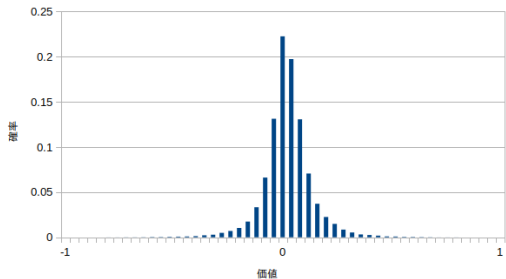
◆ 全体で R+545.8 の性能向上

- 評価精度向上で R+468.1
- 分布を考慮する探索によって R+77.7

^{iv}<https://github.com/gikou-official>

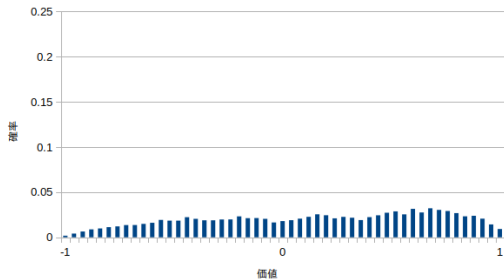
^v<http://www.uuunuuun.com/english>

出力例 1



◆ 初期局面では 0 付近に高い確率を示す

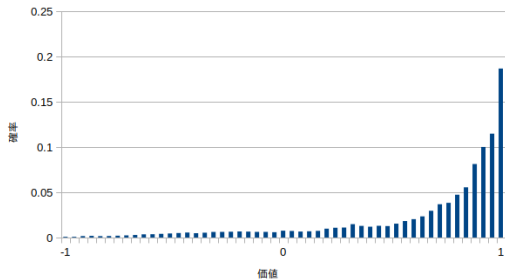
出力例 2



◆ 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