Alpha Zero

モンテカルロ木探索（MCTS）

- すべての合法手をランダムに無数に繰り返して予測を行う
- 合法手が多い場合に膨大な計算が必要（非現実）

探索と活用のジレンマ

- 活用・・・今の情報で一番得が大きそうな行動
- 探索・・・まだ、分からない新規の選択肢を選ぶ行動
- 活用を重視すると得は大きいが本当はもっと良い選択肢がある可能性
- 今 vs 将来の可能性

$$\mathrm{UCB1}(i) = \bar{X}_i + \sqrt{\frac{2\ln n}{n_i}}$$

- $\bar{X}_i$：腕 i の平均報酬
- n：全アームの **総試行回数**
- n_i：腕 i を引いた回数

多腕バンディット問題（UCTアルゴリズム）

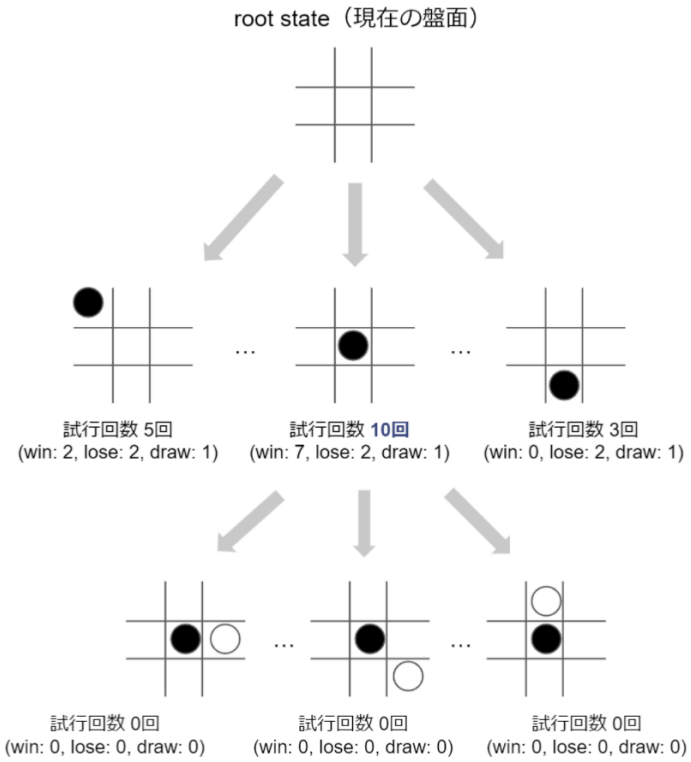
- すべての合法手についてまずは最低一回評価した後、勝利確率が高いものを優先的に深堀していく
- 勝率実績が高い（活用）が、あまり探索されていないアクションを優先的に選択する

$$\text{アクション}\mathbf{a}\text{'の選択確率} = \frac{\text{アクション}\mathbf{a}\text{'のプレイアウト累計勝利数}}{\text{アクション}\mathbf{a}\text{'の累計試行回数}} + c\sqrt{\frac{\ln \text{累計試行回数}}{\text{アクション}\mathbf{a}\text{'の累計プレイアウト回数}}}$$

- 試行回数（プレイアウト回数）のもっとも多いアクションを最善手とする。

PUTCモンテカルロ木探索

- N手先読みの要素が追加
- 評価しなければいけない盤面が増える
- 試行回数が一定回数に到達した盤面のみ、さらに次の盤面を展開



Alpha Zero

- 過去のMCTSの履歴をうまく利用して筋の悪い手を足切りする（盤面の記憶が必要）
- 盤面ごとの試行回数を記憶するのは大規模かつほんの少し盤面が異なると使えなくなる
- AlphaZeroでは過去のモンテカルロ木探索の結果を近似（再現）するようなニューラルネットワークを構築。MCTSの結果をもとにNNを訓練することで、有望な手の評価ができる
- ニューラルネット方策 P(s, a)をUCTスコアに組み込む

$$U(s,a) = c_{\mathrm{puct}} P(s,a) \frac{\sqrt{\sum_b N(s,b)}}{1 + N(s,a)}$$

- Alpha Zeroではある盤面Sを入力としその盤面Sが最終的に勝利したか敗北したかをラベルとする教師あり学習によってニューラルネットワークを訓練し、盤面の評価関数V(s)とする