AlphaZeroを用いた京都将棋AIの作成と評価

４年１４組４番石元稜

目次

目次

京都将棋の概要

- ５x５マスの盤面
- 駒は「玉」「香と」「銀角」「金桂」「飛歩」の５種。

「香と」…表「香」，裏「と」

「銀角」…表「銀」，裏「角」

「金桂」…表「金」，裏「桂」

「飛歩」…表「飛車」，裏「歩」

と表裏一体になっている。

- 駒の初期配置は右図

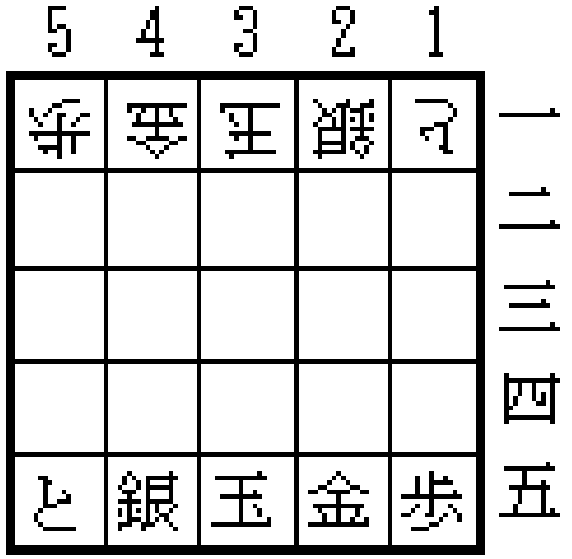


図１.京都将棋の初期配置

京都将棋のルール

基本的なルール，駒の動きは将棋と同じ。

違う点は以下の通り

・駒は１手動くごとに裏返す

・とった駒は裏表どちらでも打ってよい

・二歩はなし、千日手は引き分け。

目次

「AlphaGo」と「AlphaGo Zero」と「AlphaZero」の違い

- AlphaGo
  ハンデなしでプロに勝った初めてのコンピュータ囲碁プログラム。

- AlphaGo Zero
  プロ棋士の学習データを使わず自己対戦のみで学習し、AlphaGoに100勝0敗。

- AlphaZero
  「AlphaGo Zero」の改造バージョン。
  囲碁だけでなくチェスや将棋も学習できるようになった。

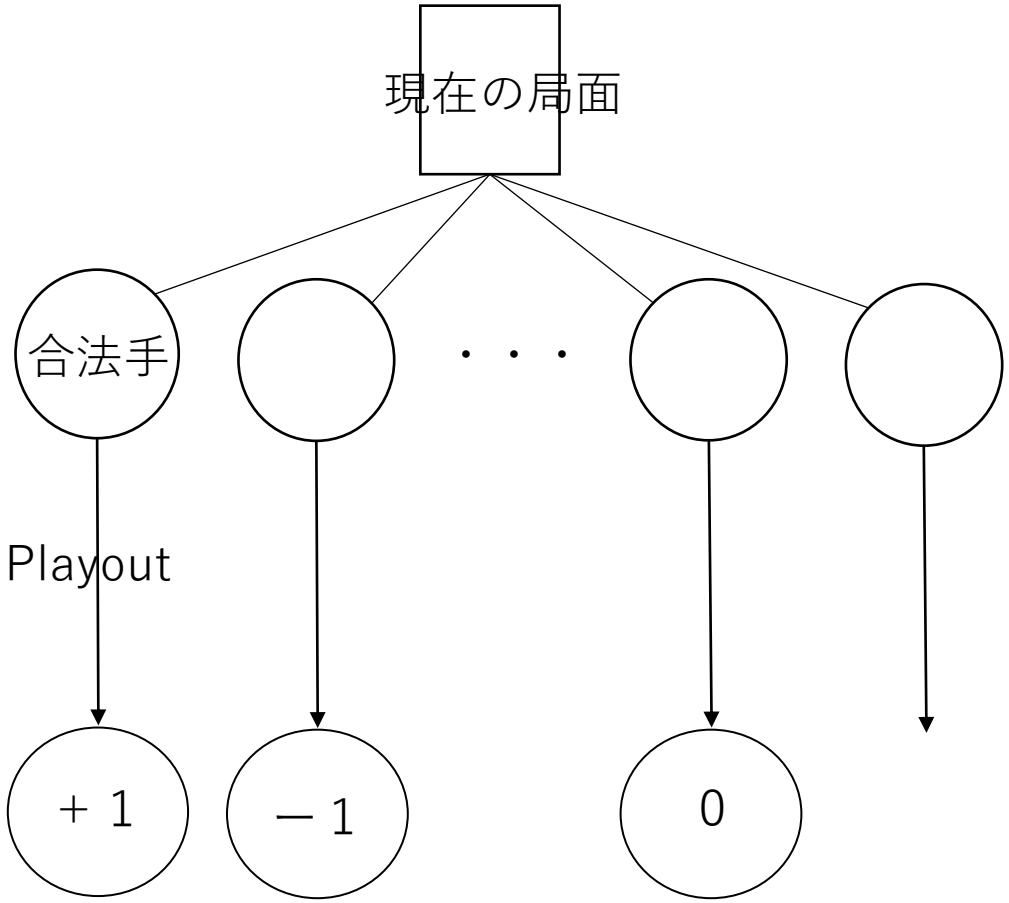
モンテカルロ木探索の前に、

- **原始モンテカルロ探索**

よりよい次の一手を探索する。

合法手（現在の局面から指せる手）を
指した局面から
何度か終局までランダムにプレイし
（playout）、
勝ち(+1)負け(-1)の合計を計算する。

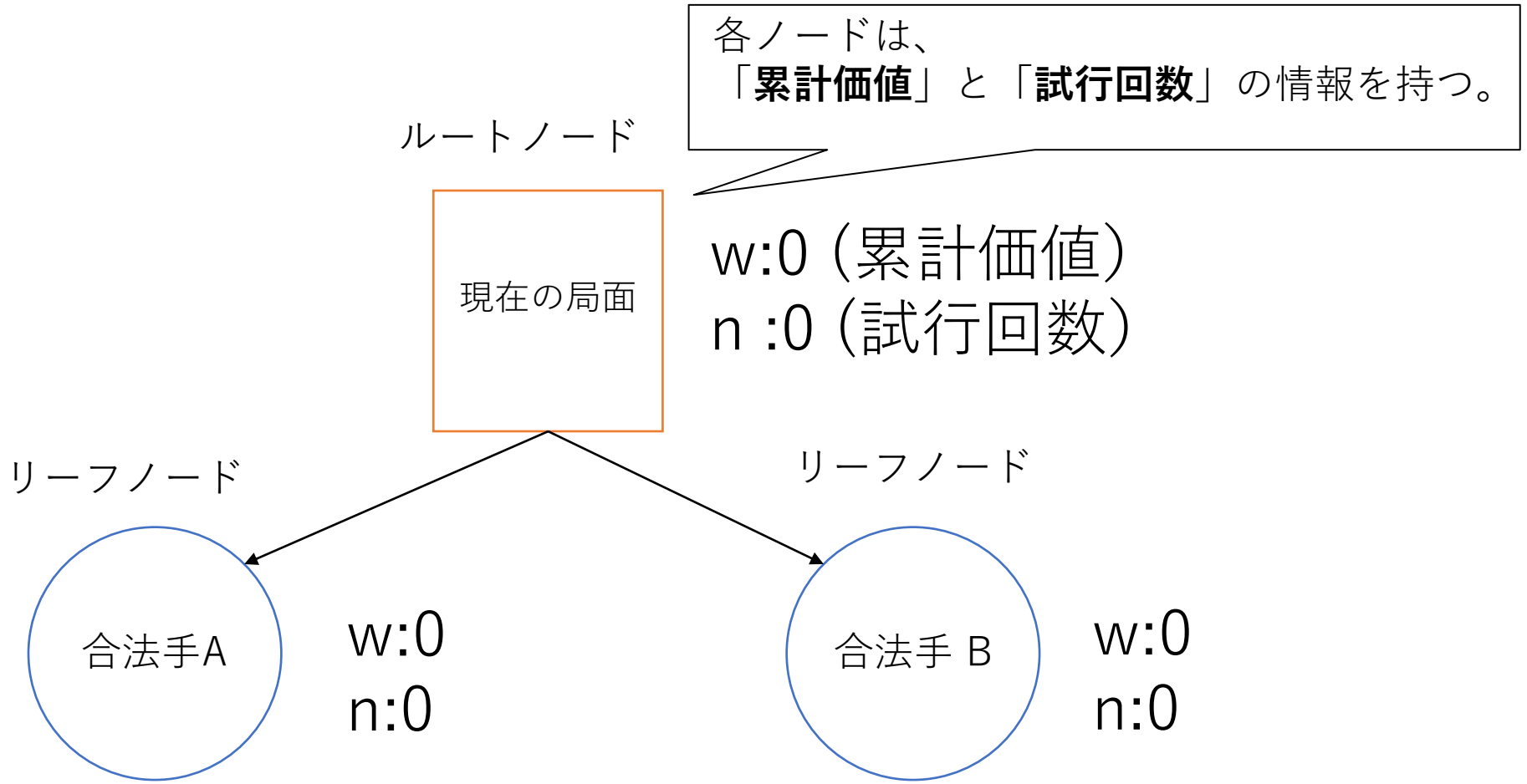
それぞれの合法手の中で
合計得点が最も大きい手を選択。



モンテカルロ木探索

モンテカルロ木探索は原始モンテカルロ探索に
「**選択**」、「**展開**」、「**評価**」、「**更新**」
の４つの操作を加えたもの。

モンテカルロ木探索ー初期状態



モンテカルロ木探索－選択

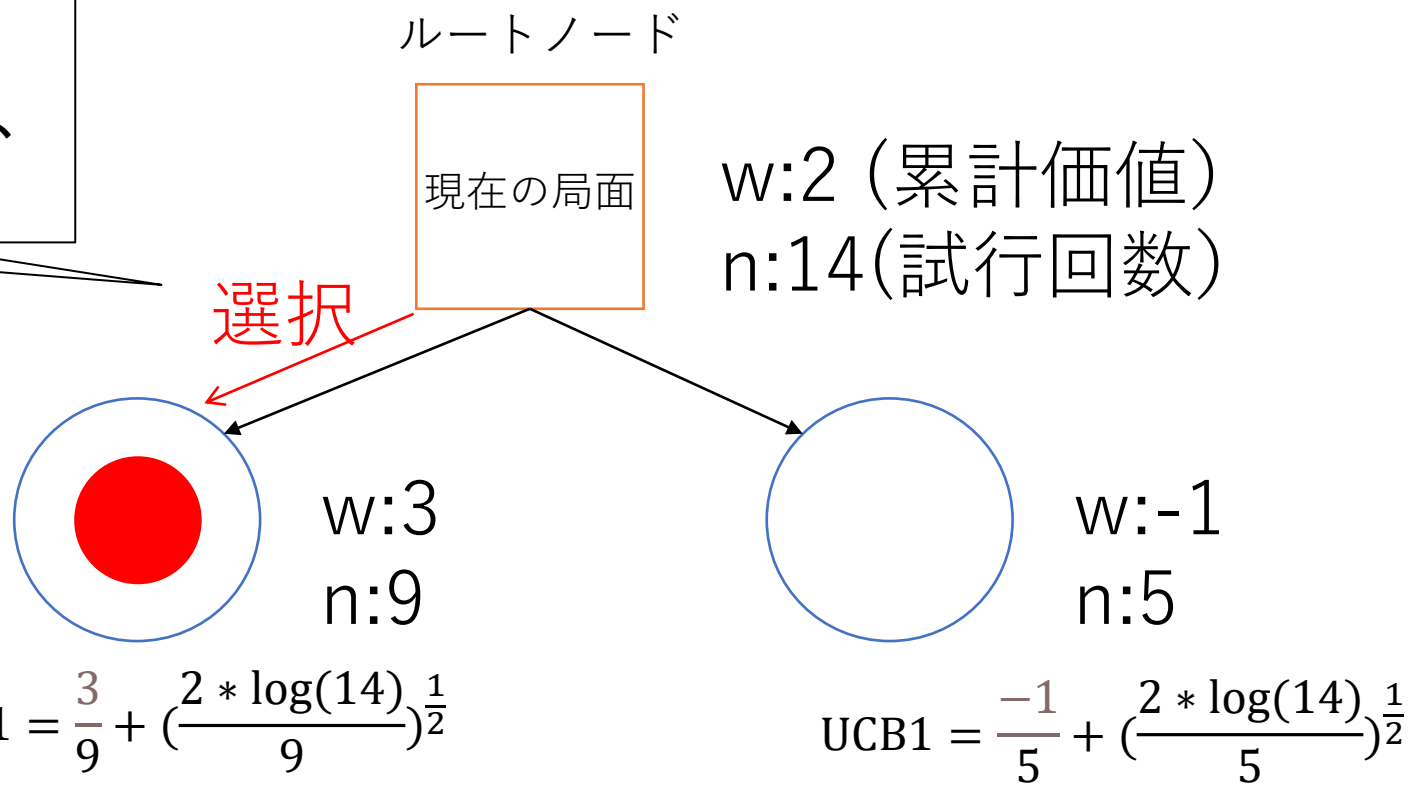
「ルートノード」から「子ノード」が存在したら「リーフノード」に到達するまで移動する。

この時、「UCB1」（バイアス＋確率）に従い、移動先を選択する。

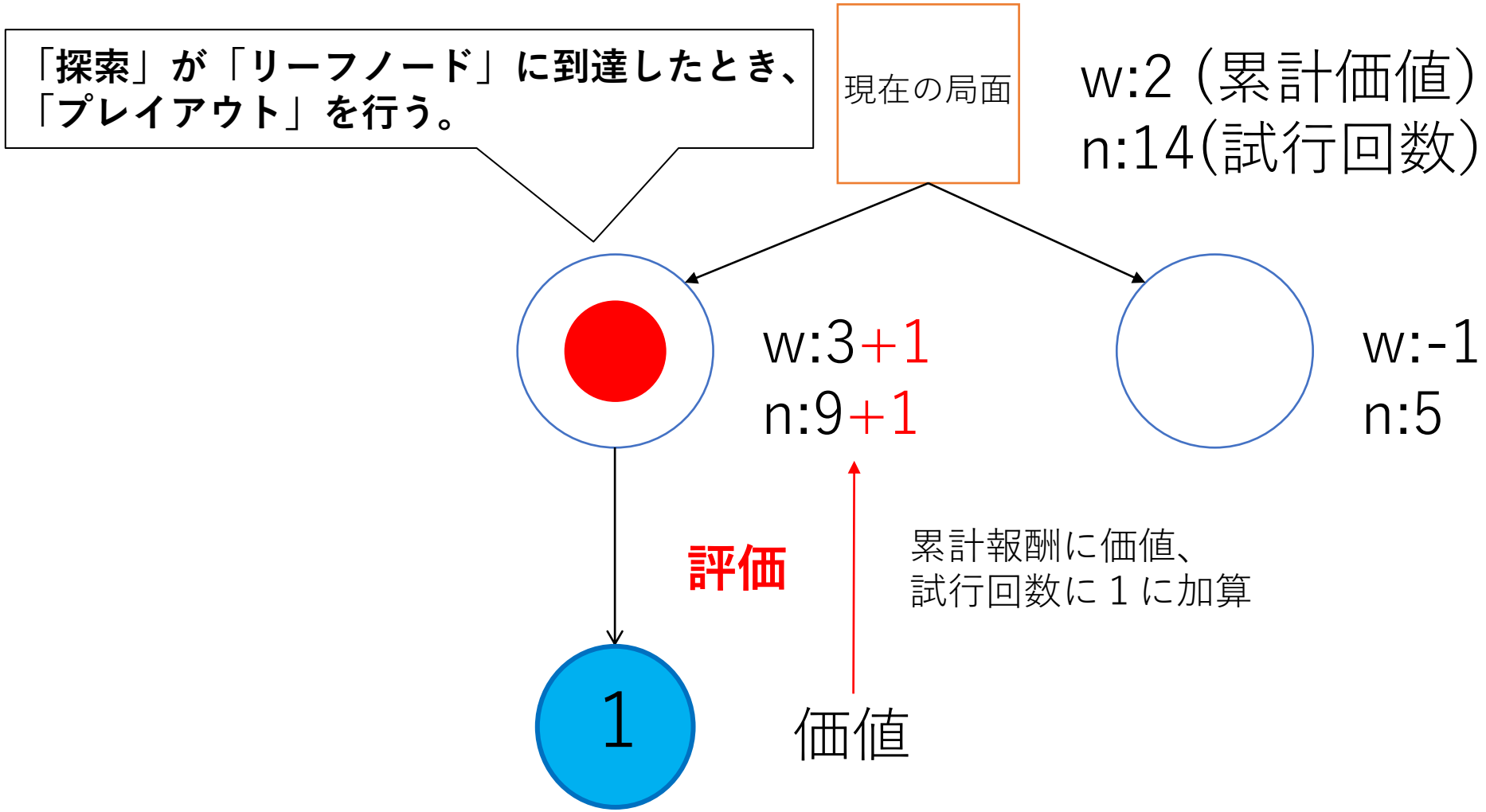
$$\mathrm{UCB1} = \frac{w}{n} + \left(\frac{2 * \log(t)}{n}\right)^{\frac{1}{2}}$$

成功率　　バイアス

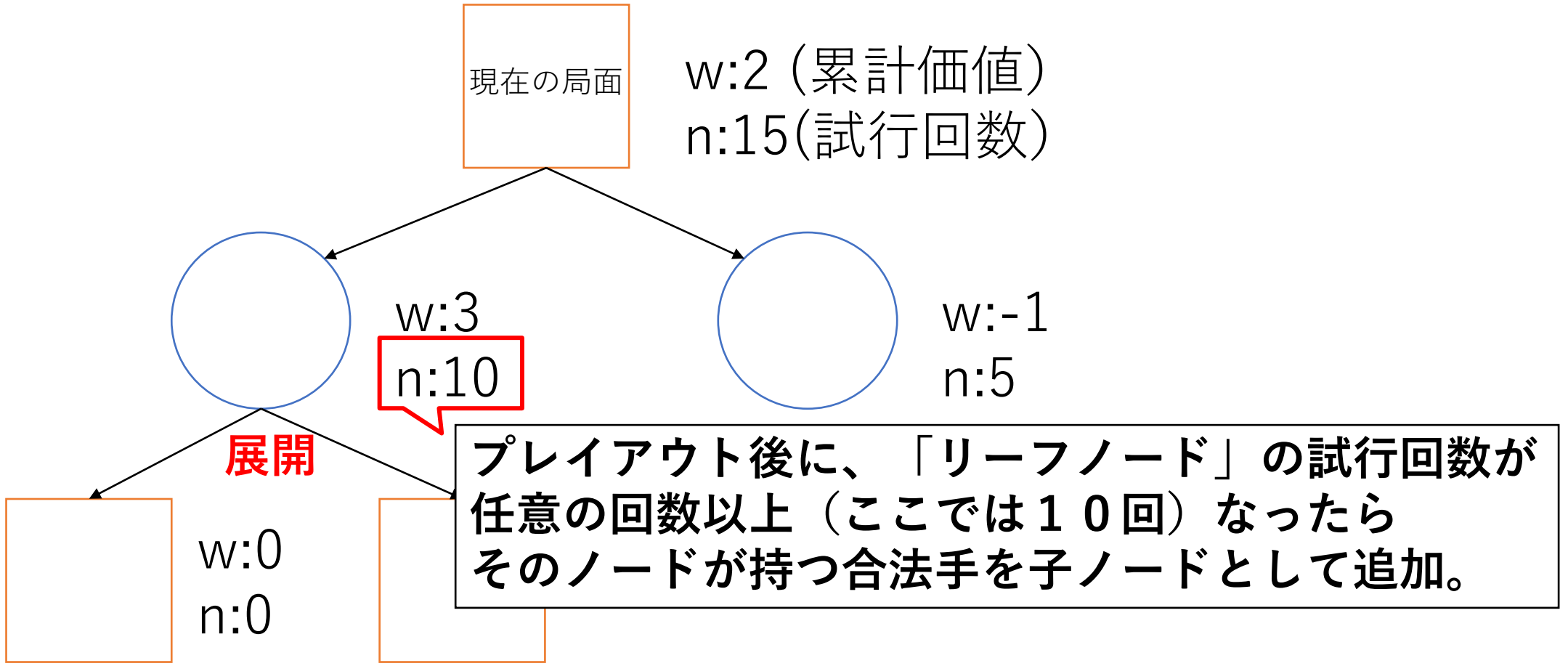
n：この行動の試行回数
w：この行動の累計価値
t：すべての行動の試行回数



モンテカルロ木探索ー評価

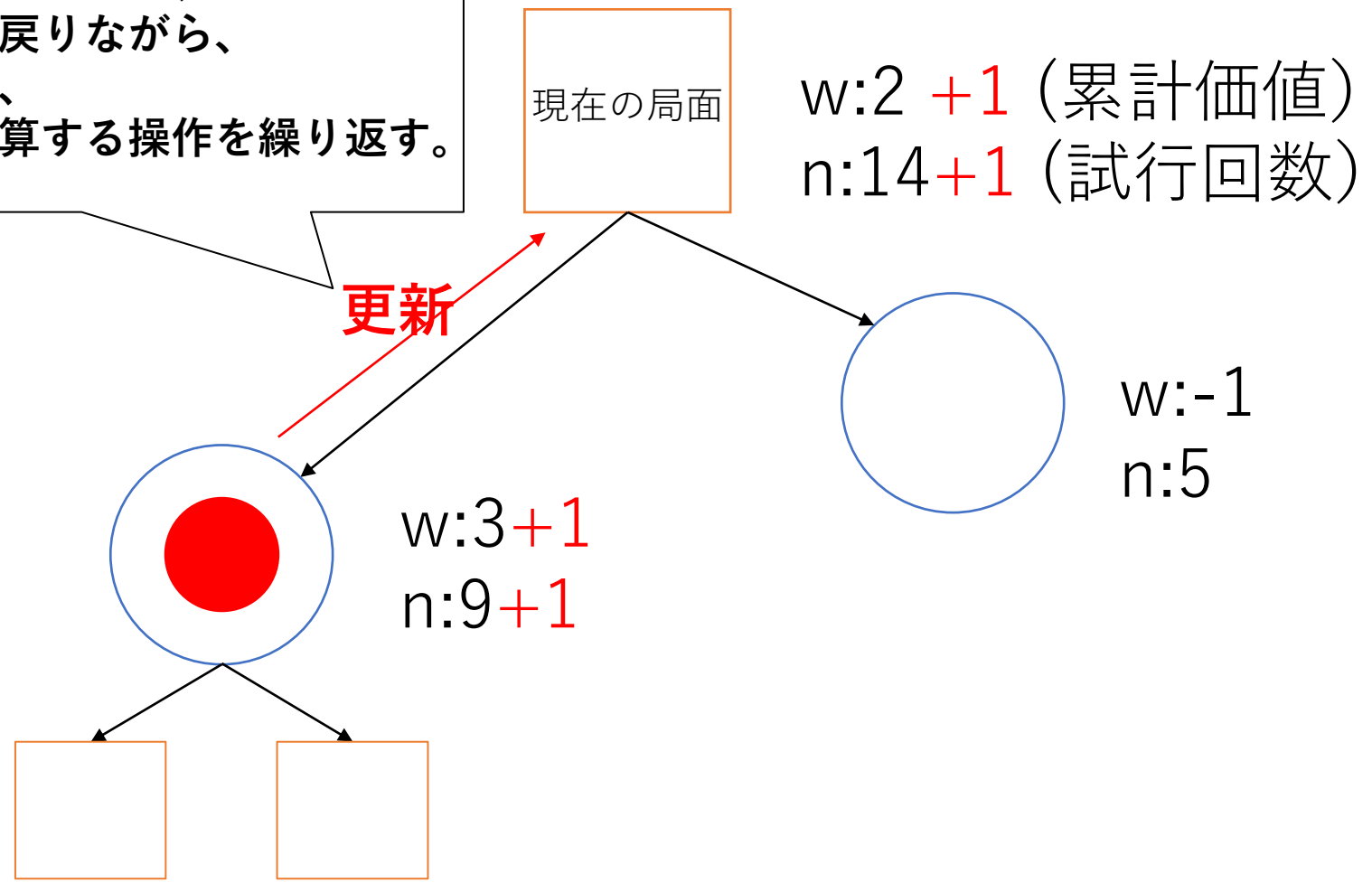


モンテカルロ木探索ー展開



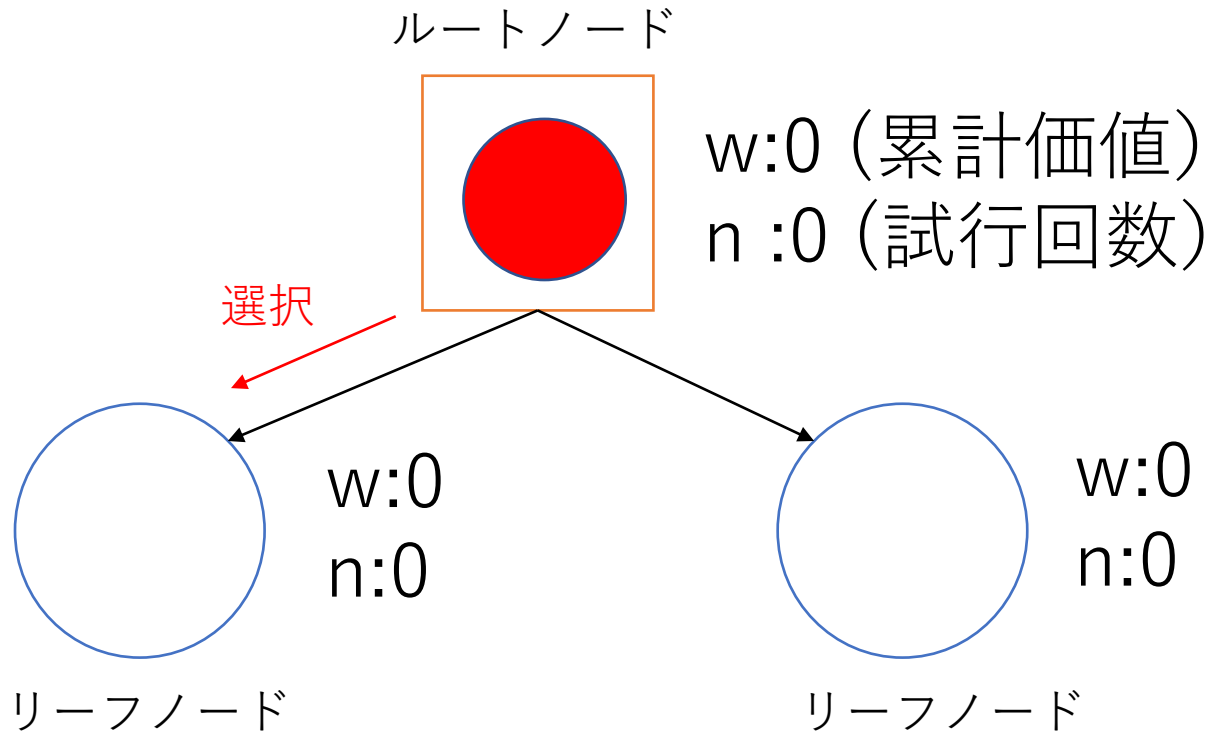
モンテカルロ木探索ー更新

「プレイアウト」が終わったら、
「ルートノード」まで戻りながら、
ノードの「累計価値」、
「試行回数」に１を換算する操作を繰り返す。



モンテカルロ木探索ーシミュレーション

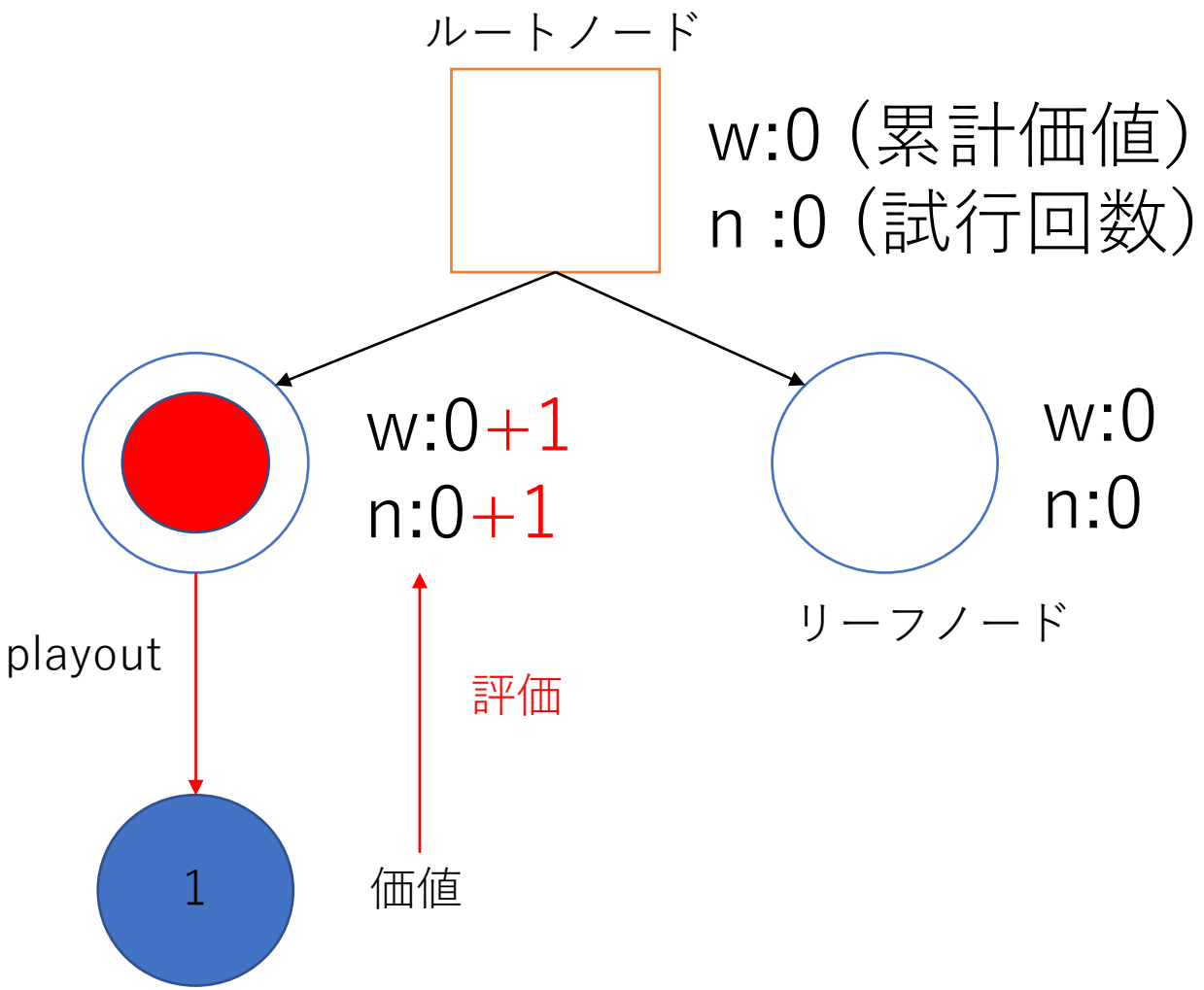
１回目



「UCB1」は試行回数が１以上じゃないと計算できないため、
試行回数が０のノードを一通り「選択」する。

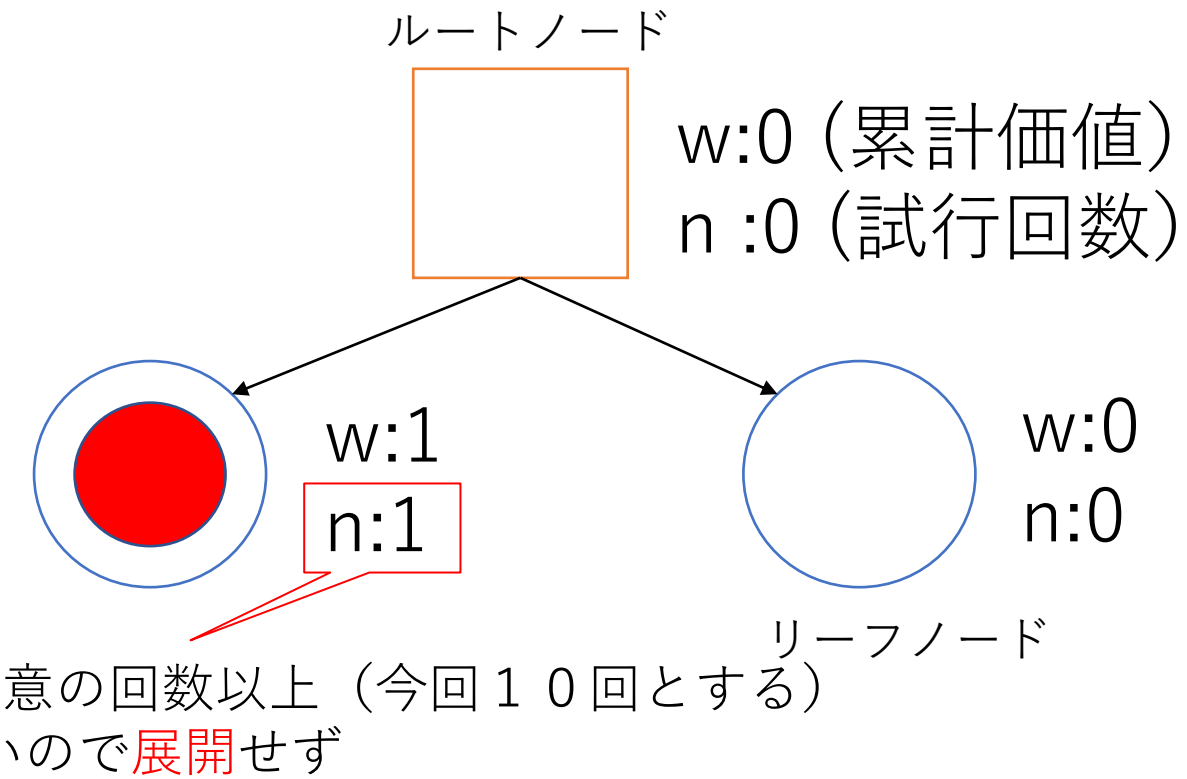
モンテカルロ木探索ーシミュレーション

１回目



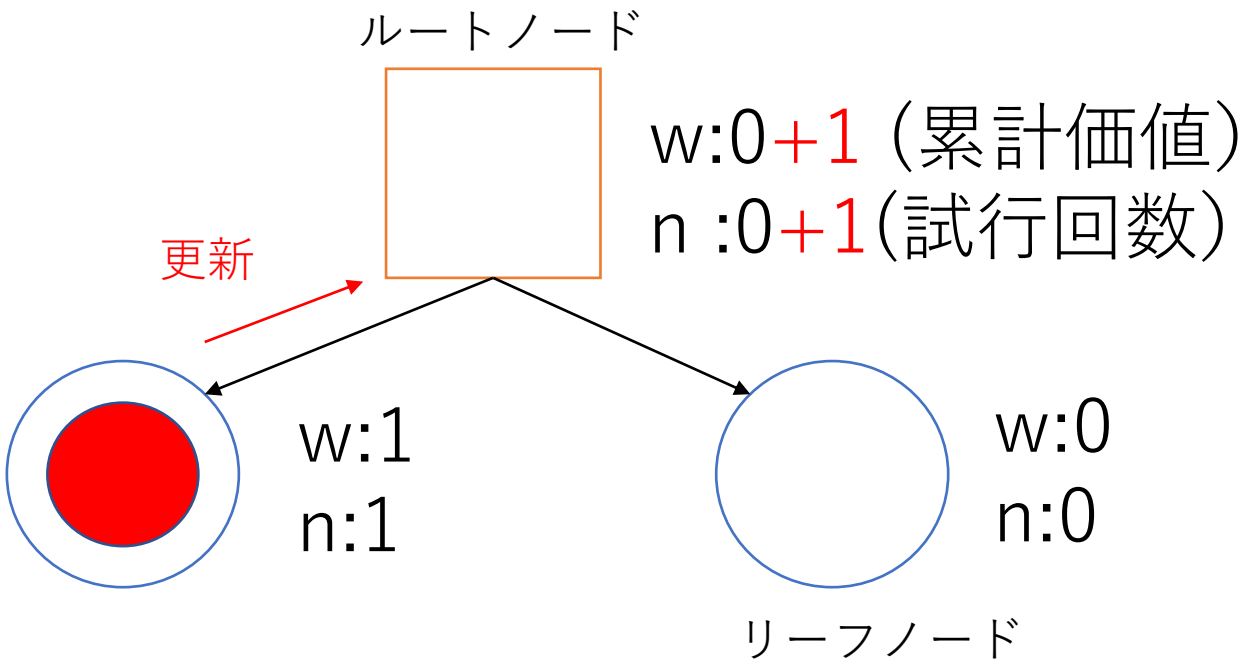
モンテカルロ木探索ーシミュレーション

１回目



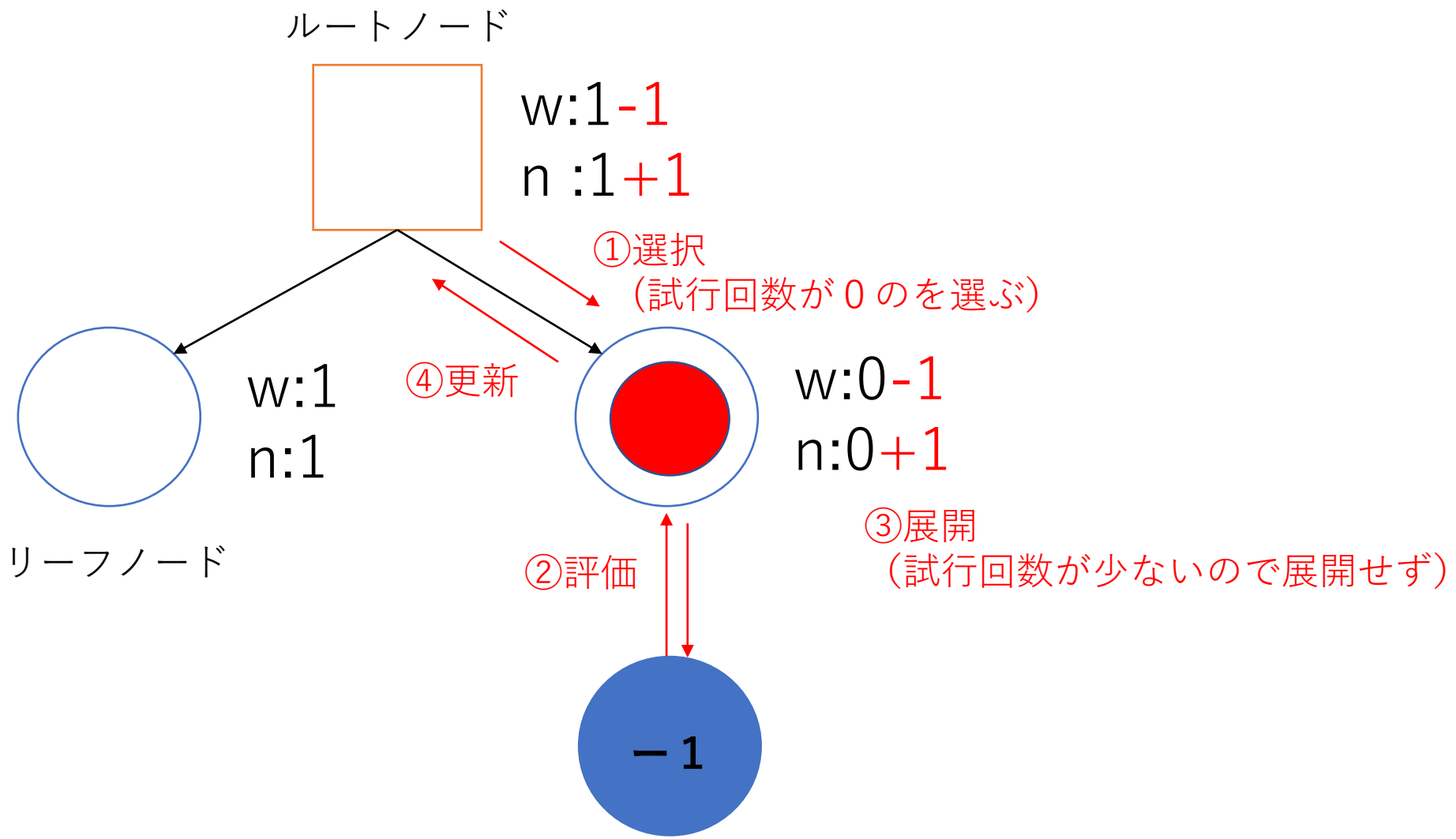
モンテカルロ木探索ーシミュレーション

１回目



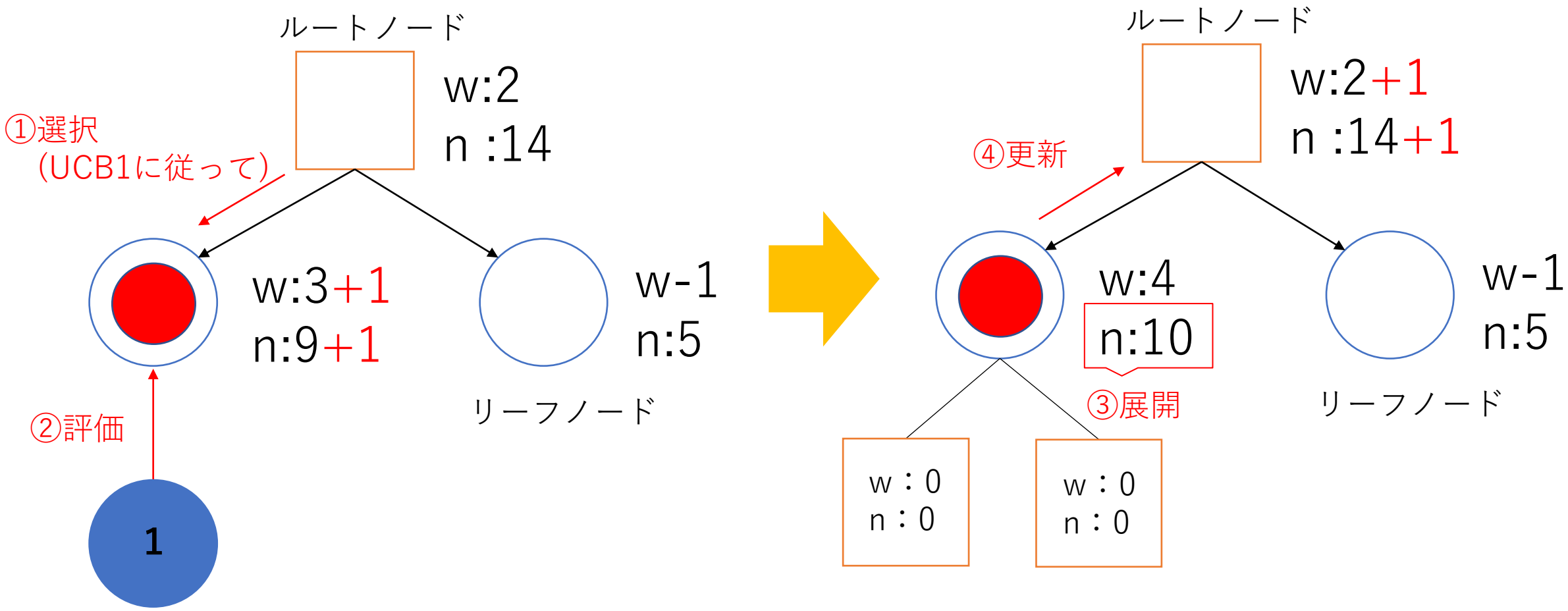
モンテカルロ木探索ーシミュレーション

２回目



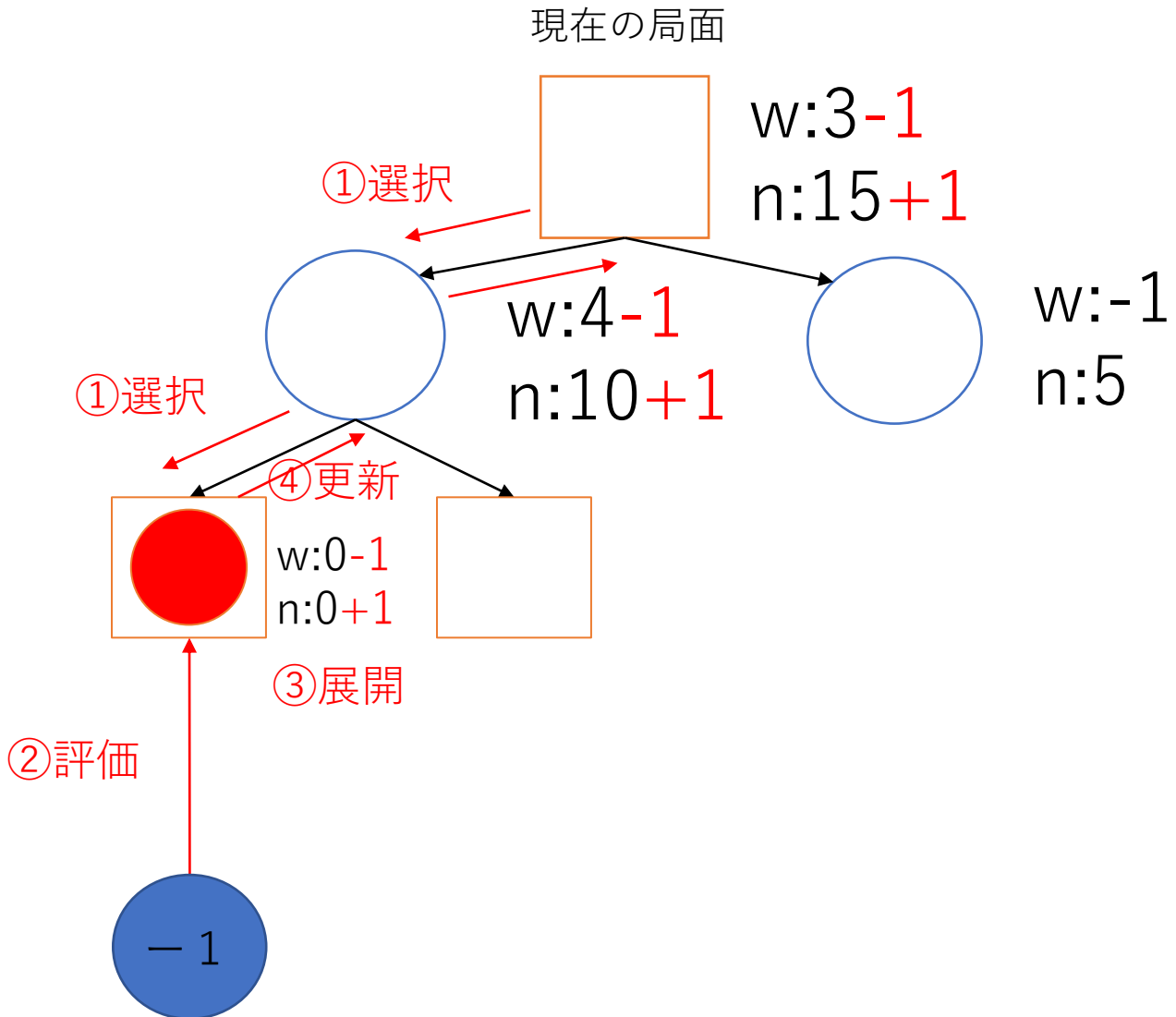
モンテカルロ木探索ーシミュレーション

１５回目



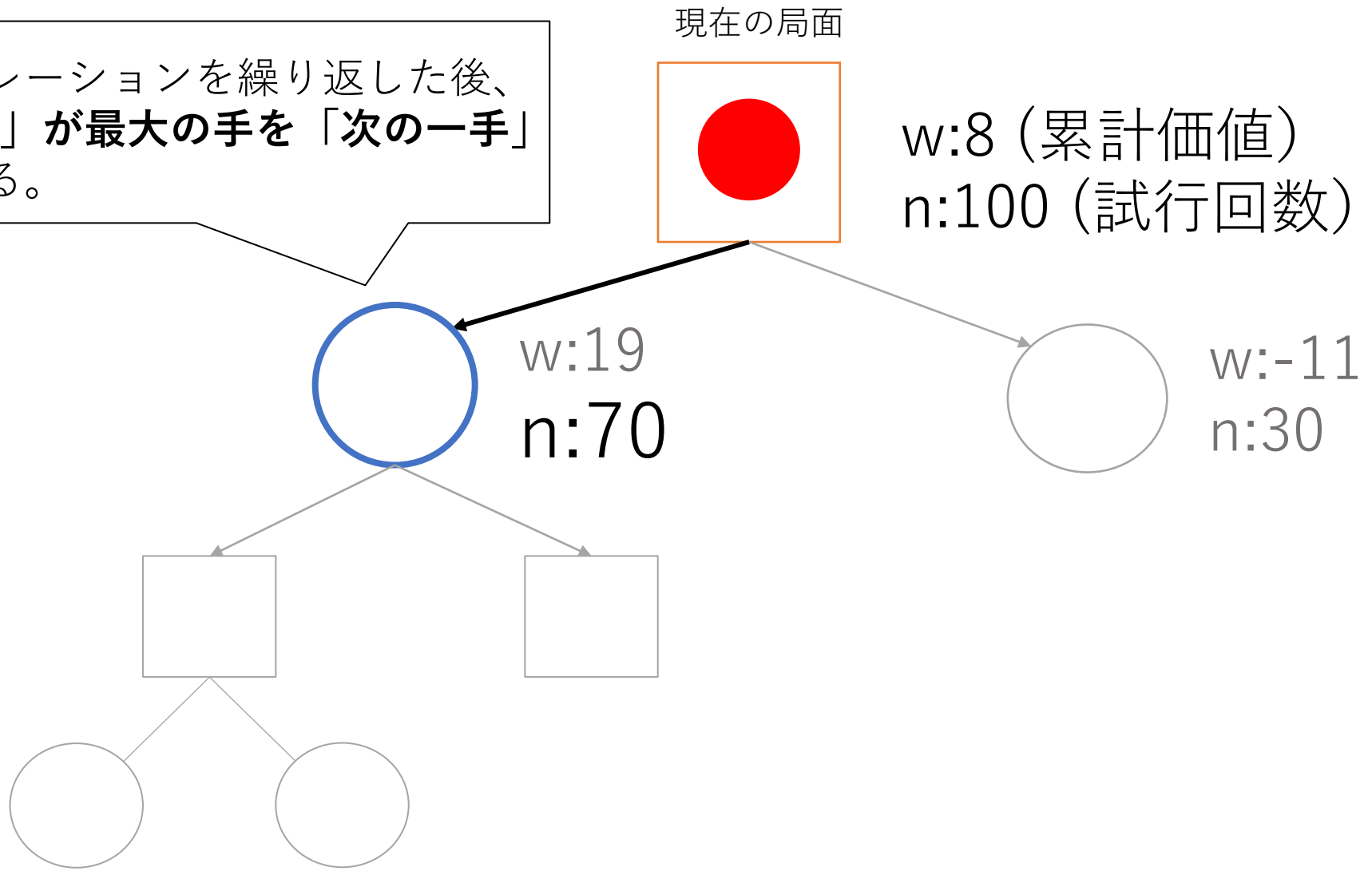
モンテカルロ木探索ーシミュレーション

１６回目

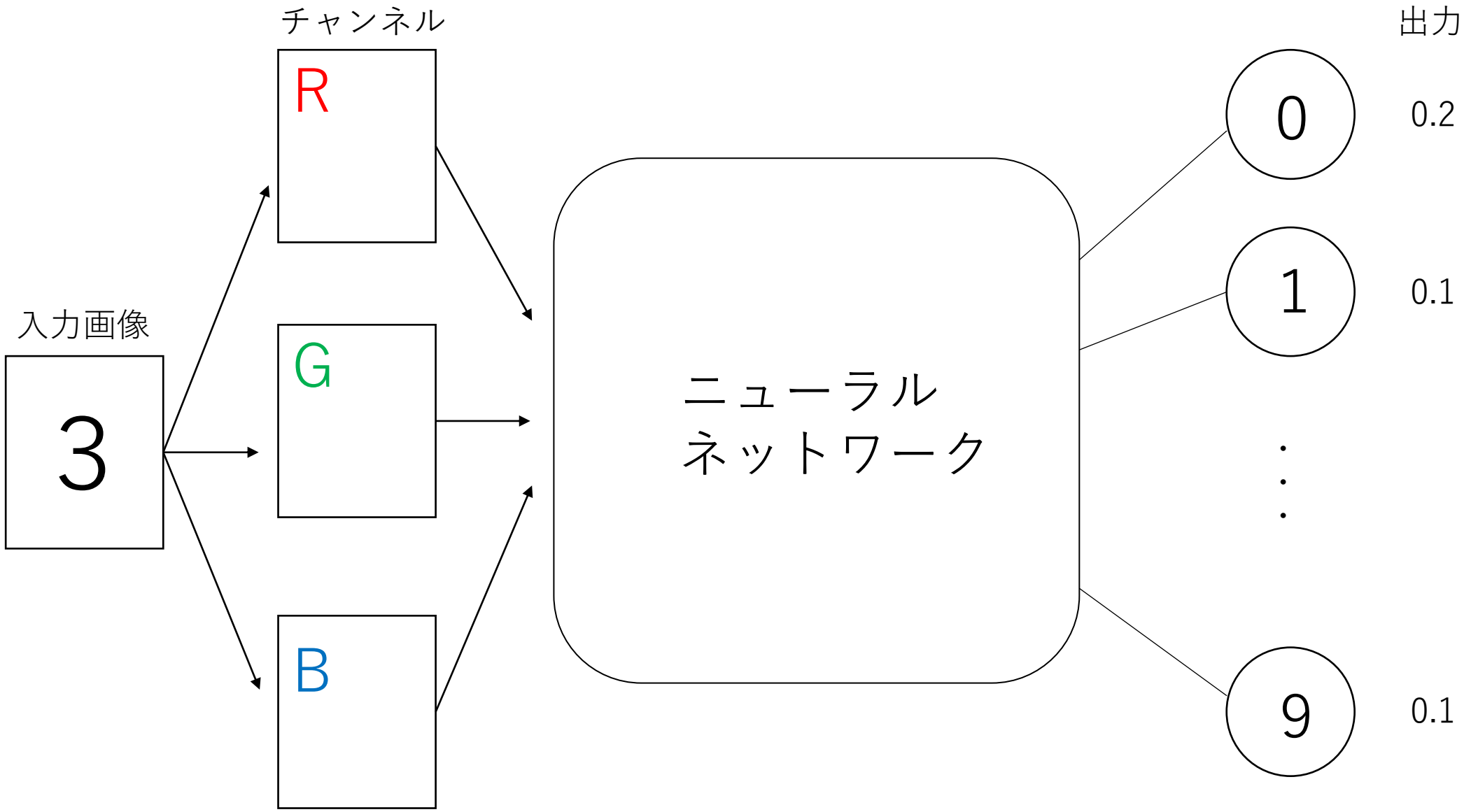


モンテカルロ木探索ー行動選択

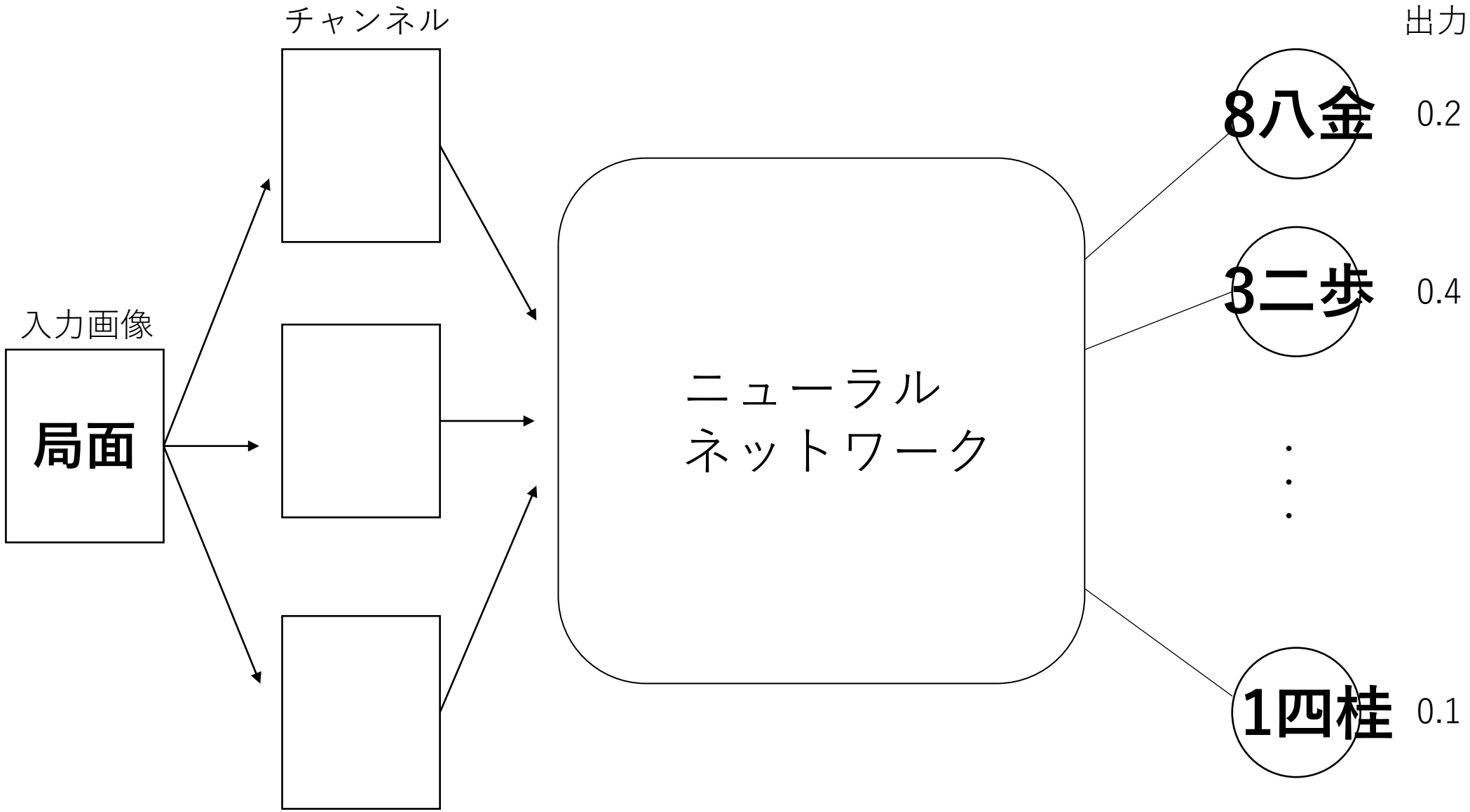
十分にシミュレーションを繰り返した後、
「**試行回数(n)**」**が最大の手を**「**次の一手**」
として選択する。



ニューラルネットワークの画像識別の例

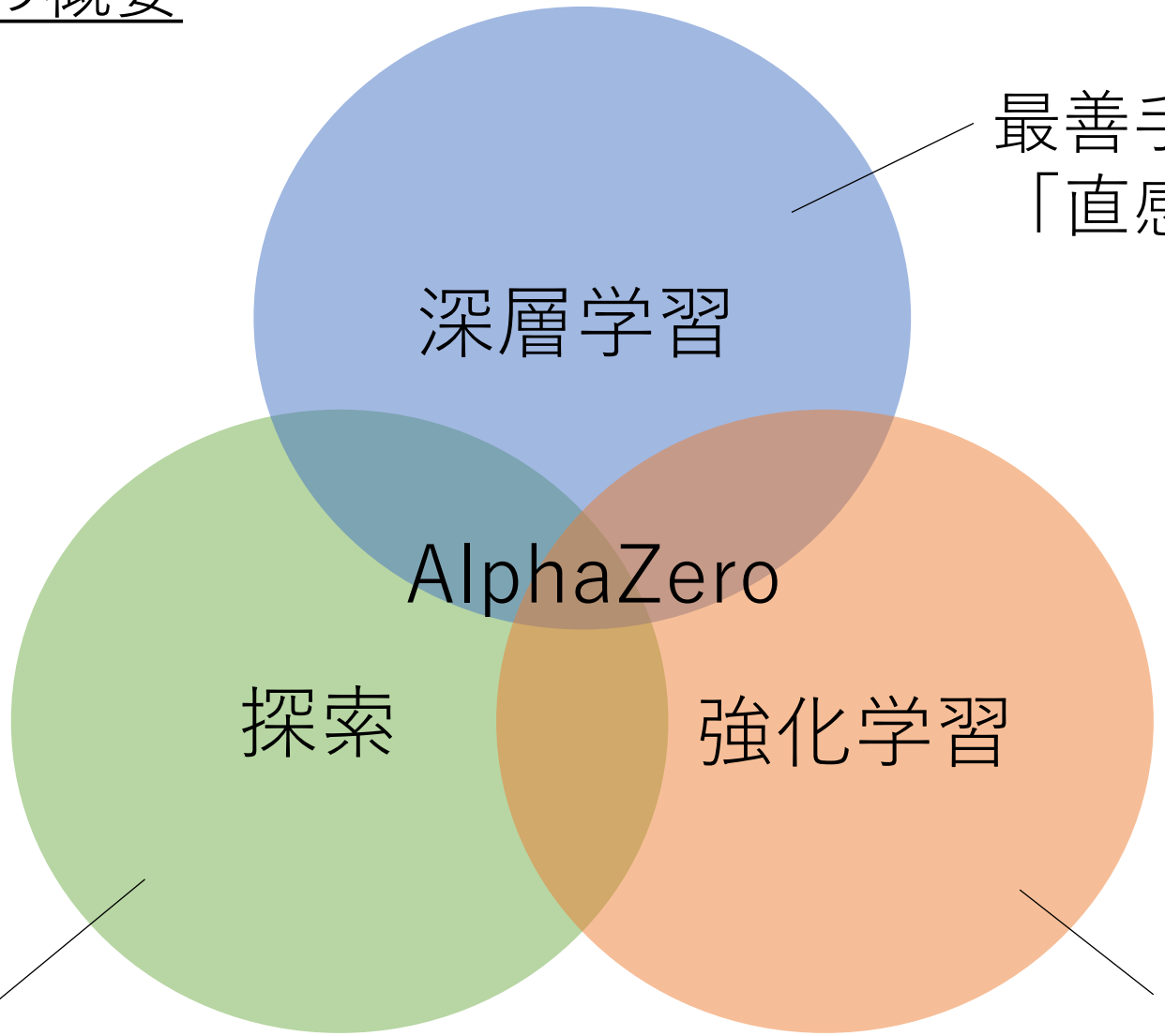


ニューラルネットワークの将棋への活用



目次

「AlphaZero」の概要

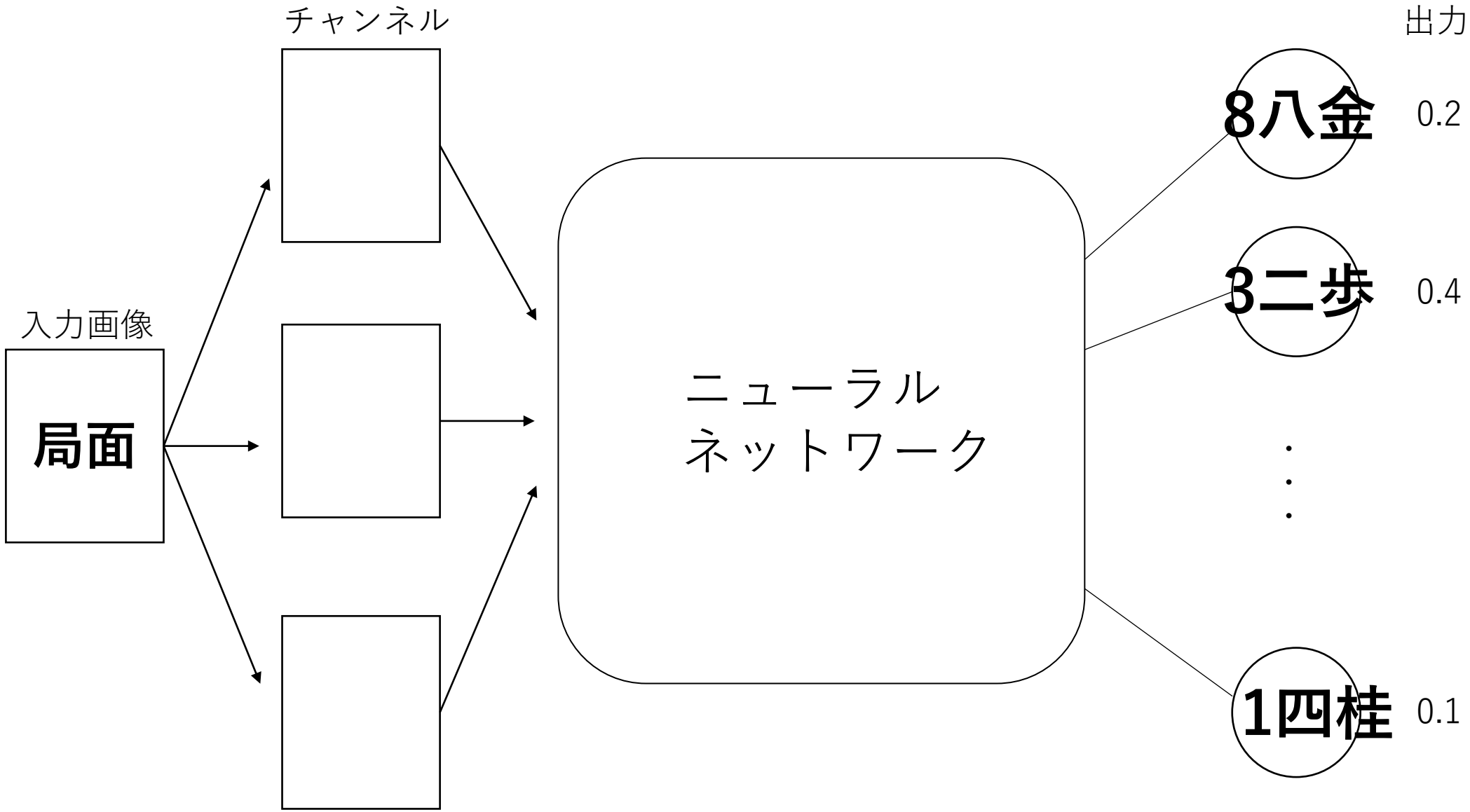


最善手を予測する
「直感」

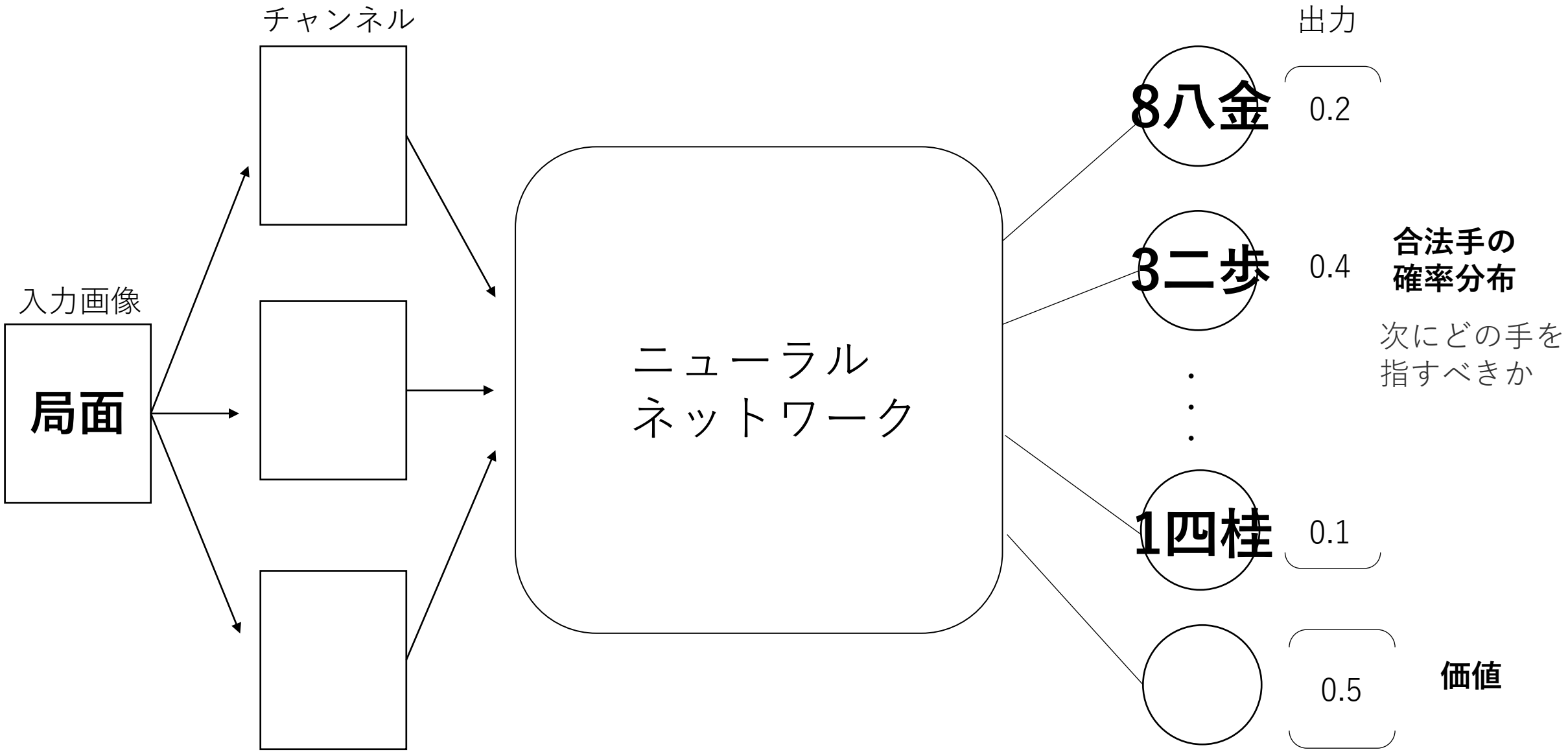
「先読みする力」

自己対戦による
「経験」

ニューラルネットワークの将棋への活用（再掲）



AlphaZeroのニューラルネットワーク



AlphaZeroのモンテカルロ木探索ー初期状態

各ノードは、
「累計価値」、「試行回数」に加え
「**それぞれの合法手の確率**」の情報を持つ。

ルートノード

現在の局面

w:0（累計価値）
n:0（試行回数）
p:0

w:0
n:0
p:0.6

リーフノード

w:0
n:0
p:0.4

リーフノード

AlphaZeroのモンテカルロ木探索ー選択

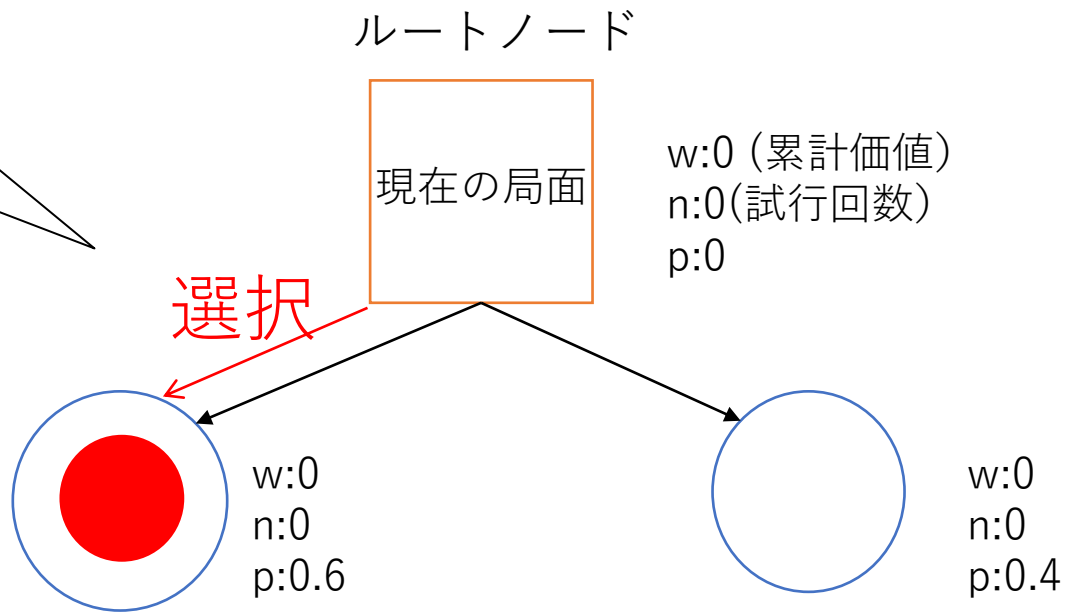
「ルートノード」から「子ノード」が存在したら
「リーフノード」に到達するまで移動する。

この時、「UCB1」（バイアス＋確率）に従い、
移動先を選択する。

$$\text{アーク評価値} = \frac{w}{n} + c_{puct} * p * \frac{\sqrt{t}}{(1+n)}$$

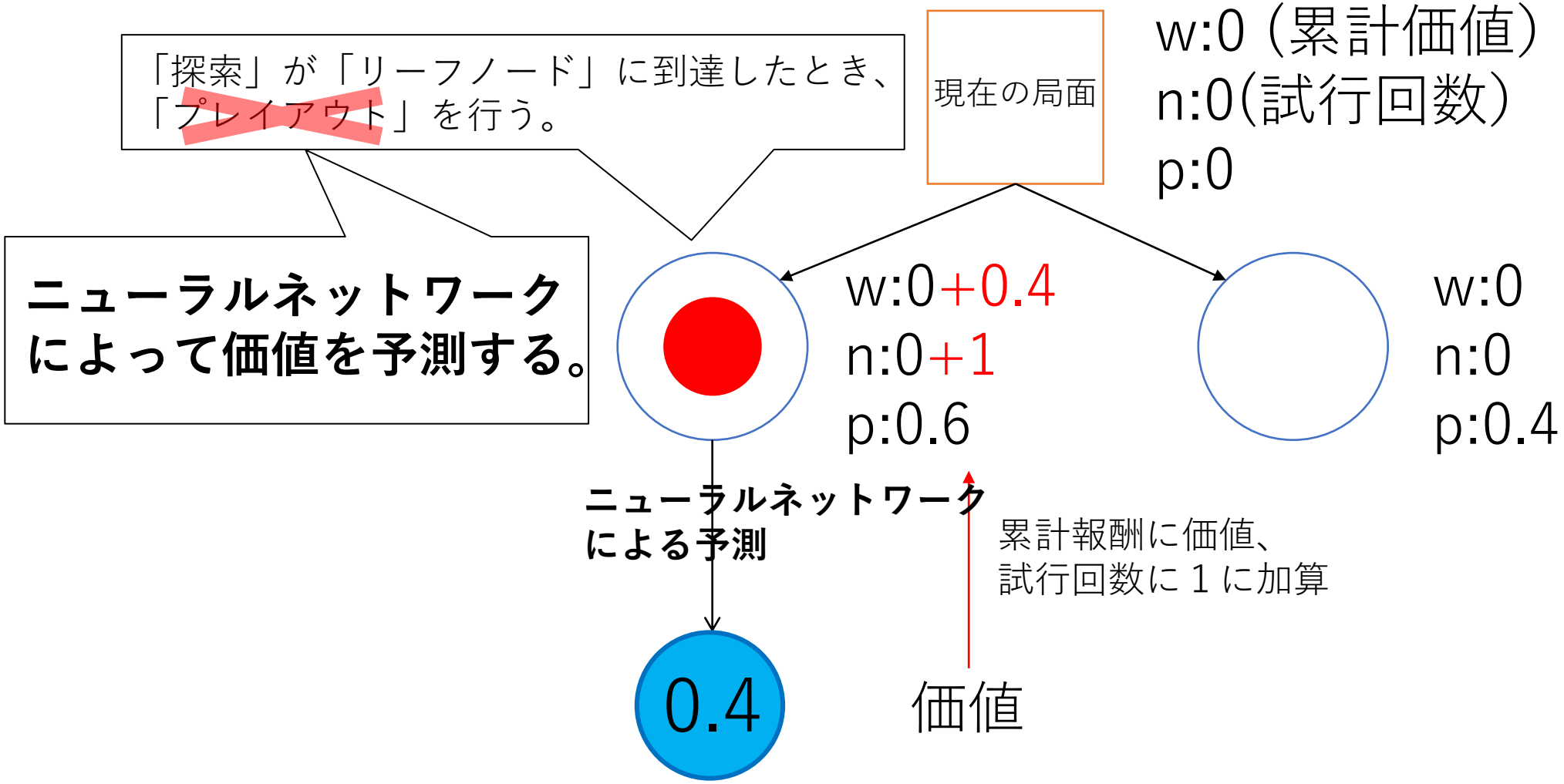
合法手の確率分布

成功率　バランス調整の定数　バイアス

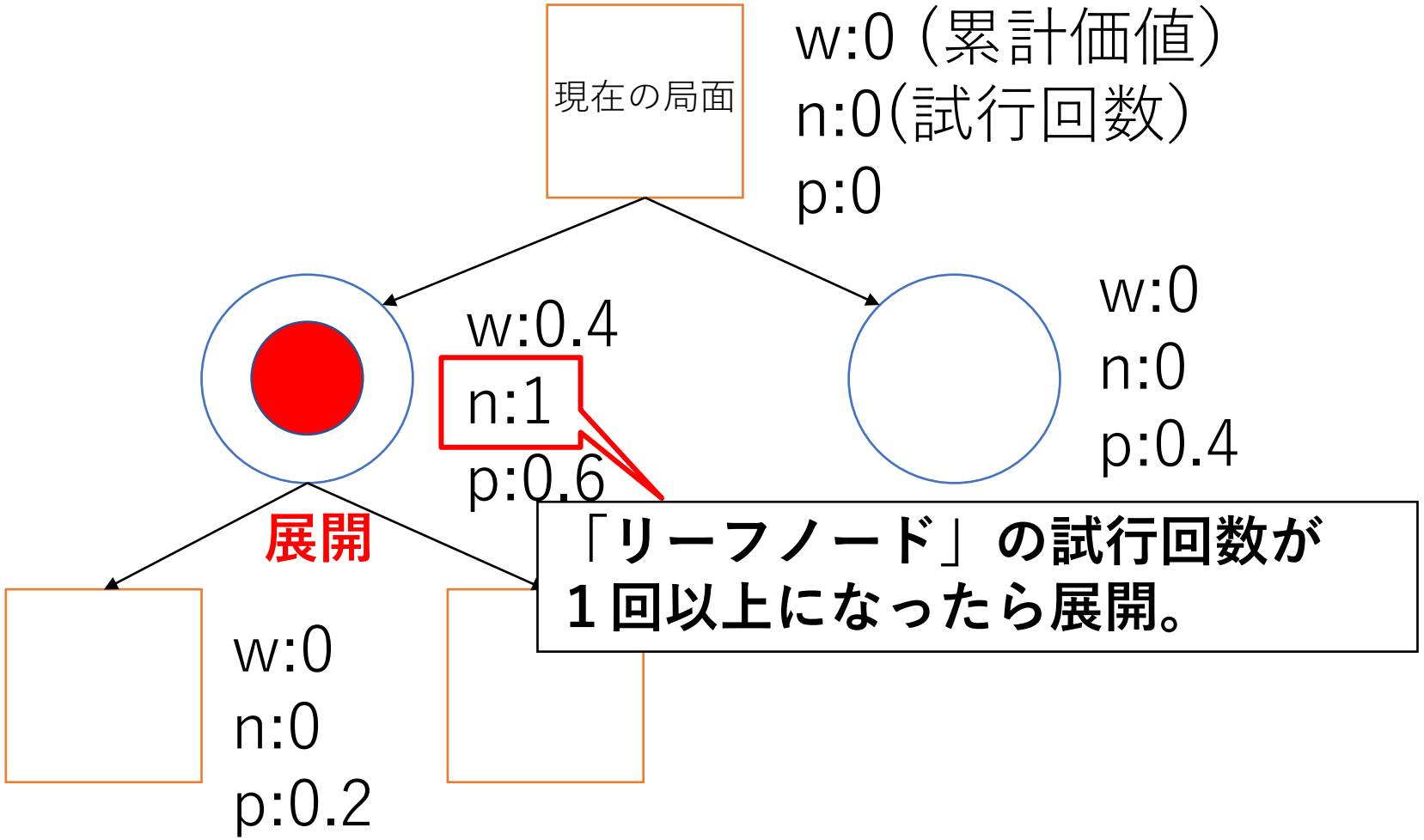


n：このノードの試行回数，　w：このノードの累計価値，
t：累計試行回数
Cpuct：「勝率」と「手の予測確率*バイアス」のバランスを調整するための定数

AlphaZeroのモンテカルロ木探索ー評価

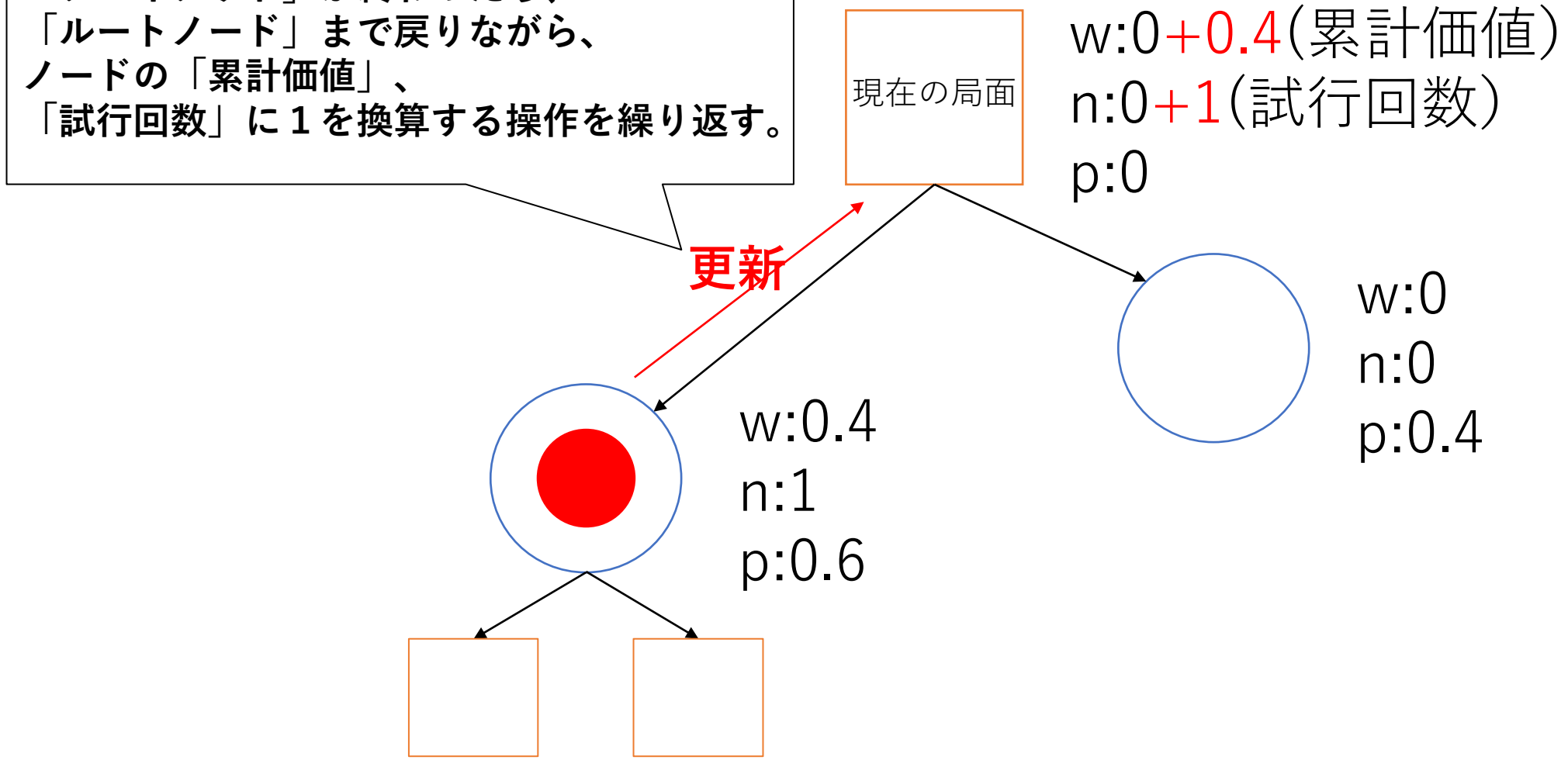


AlphaZeroのモンテカルロ木探索ー展開

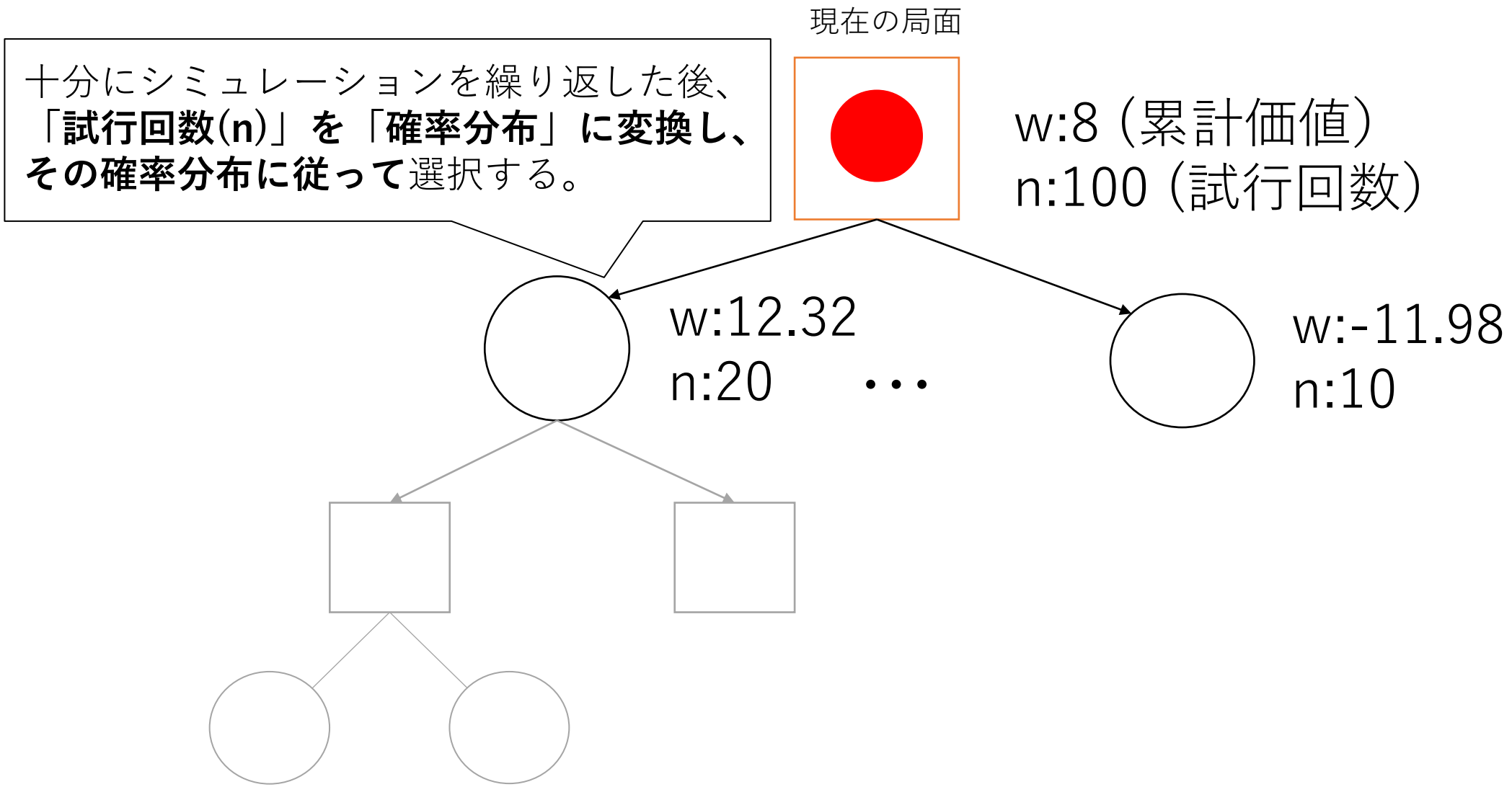


AlphaZeroのモンテカルロ木探索ー更新

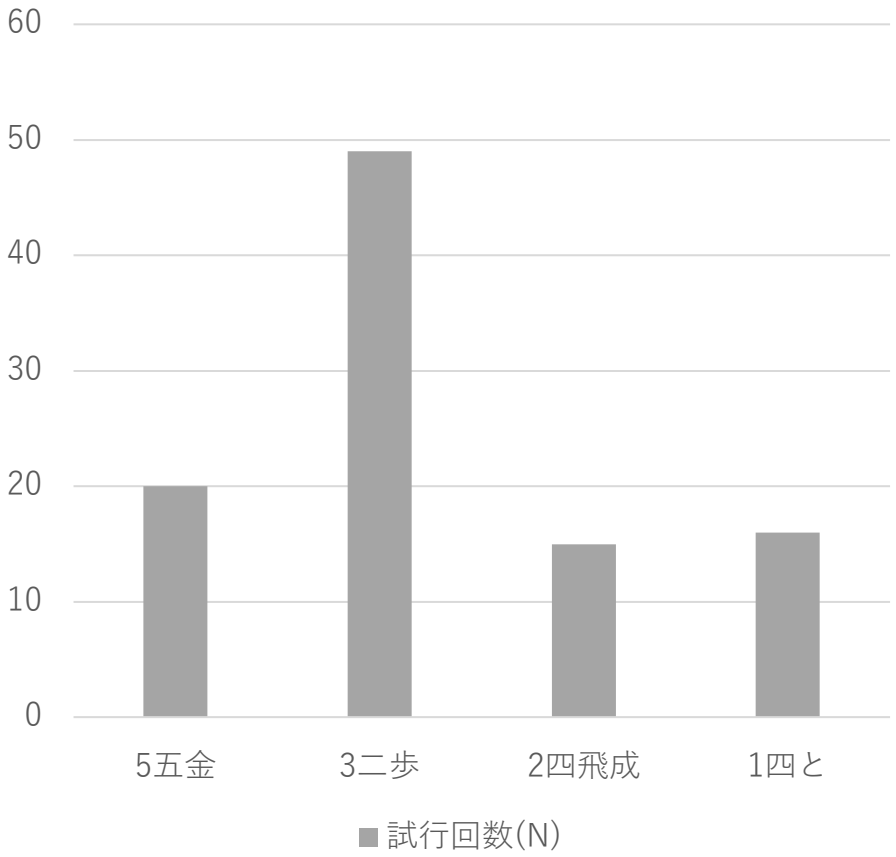
「プレイアウト」が終わったら、
「ルートノード」まで戻りながら、
ノードの「累計価値」、
「試行回数」に１を換算する操作を繰り返す。

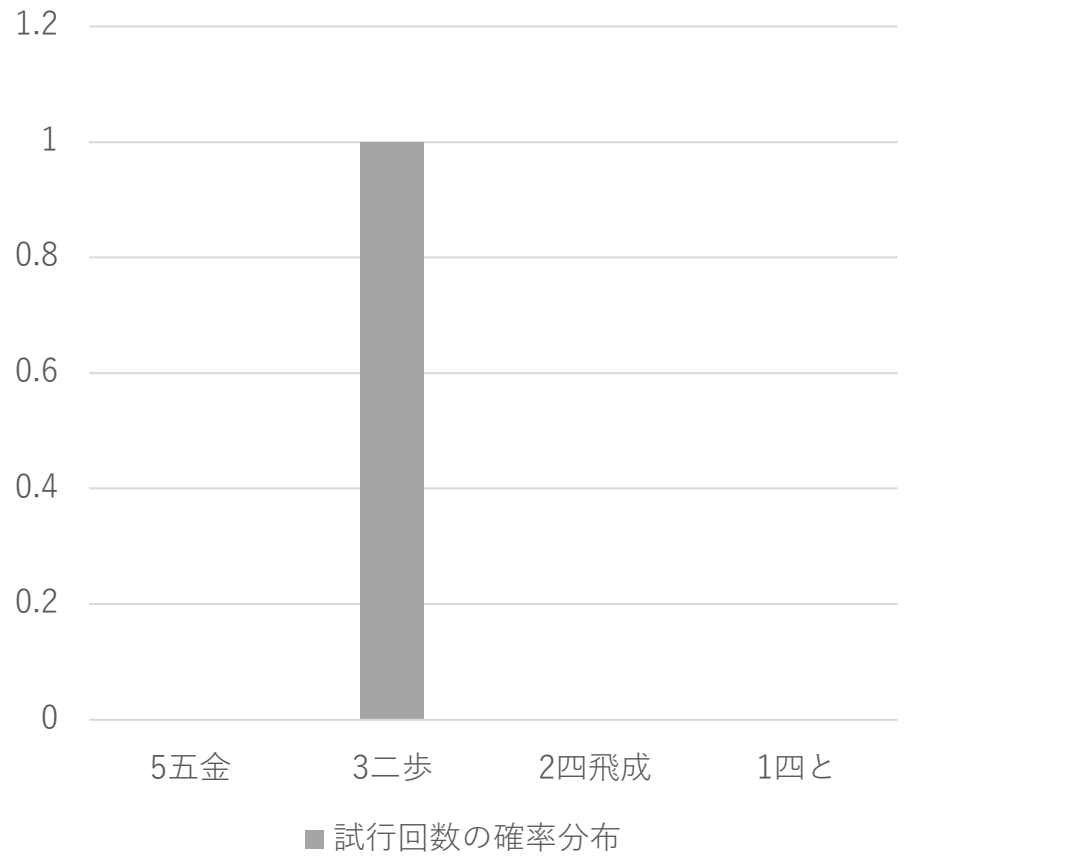


AlphaZeroのモンテカルロ木探索ー行動選択



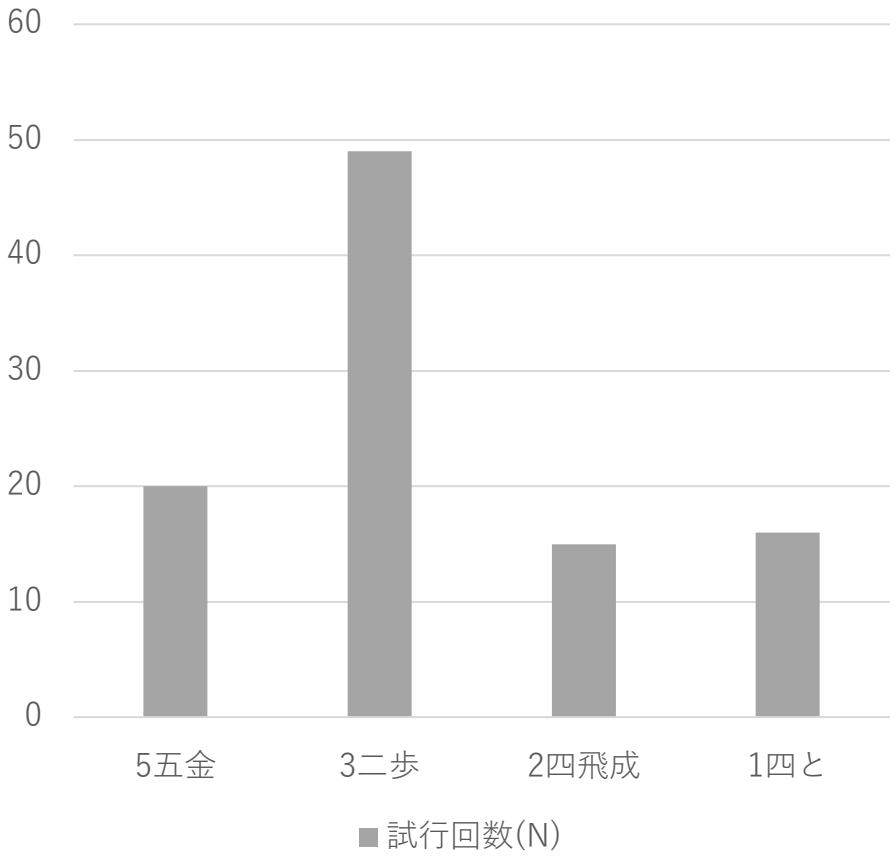
AlphaZeroのモンテカルロ木探索ー行動選択（本番時）

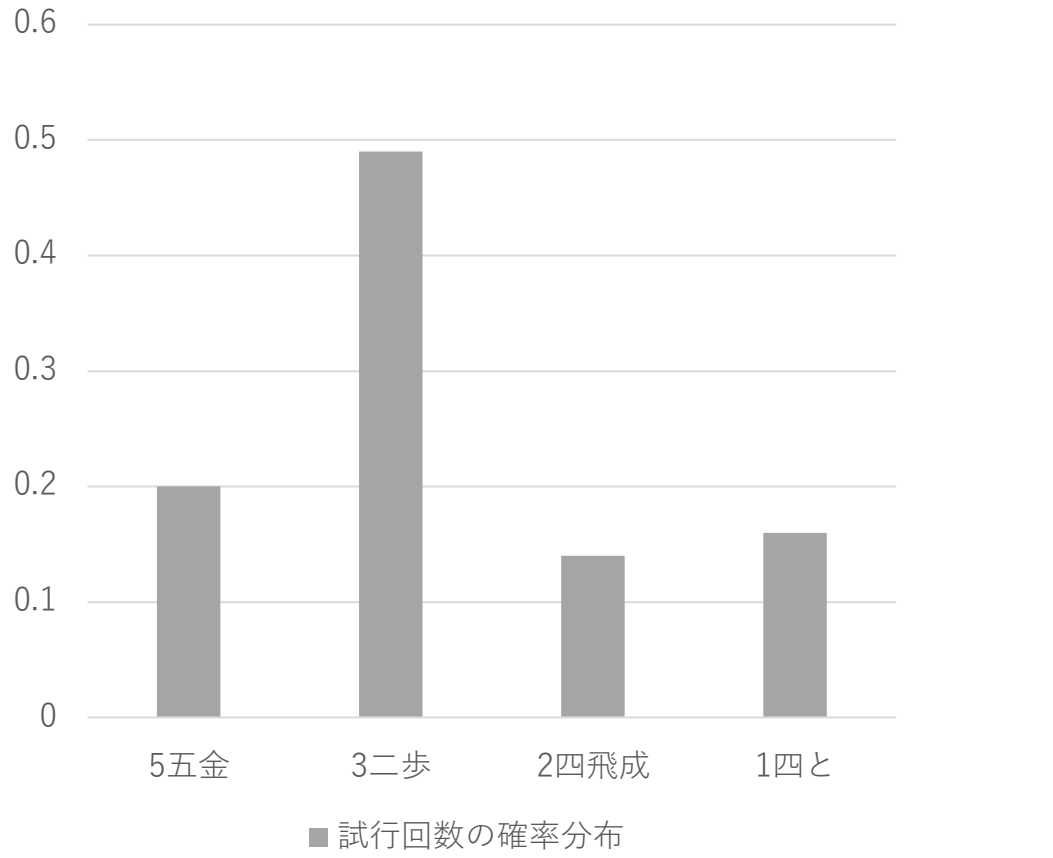




この確率分布に従って１手を選ぶ

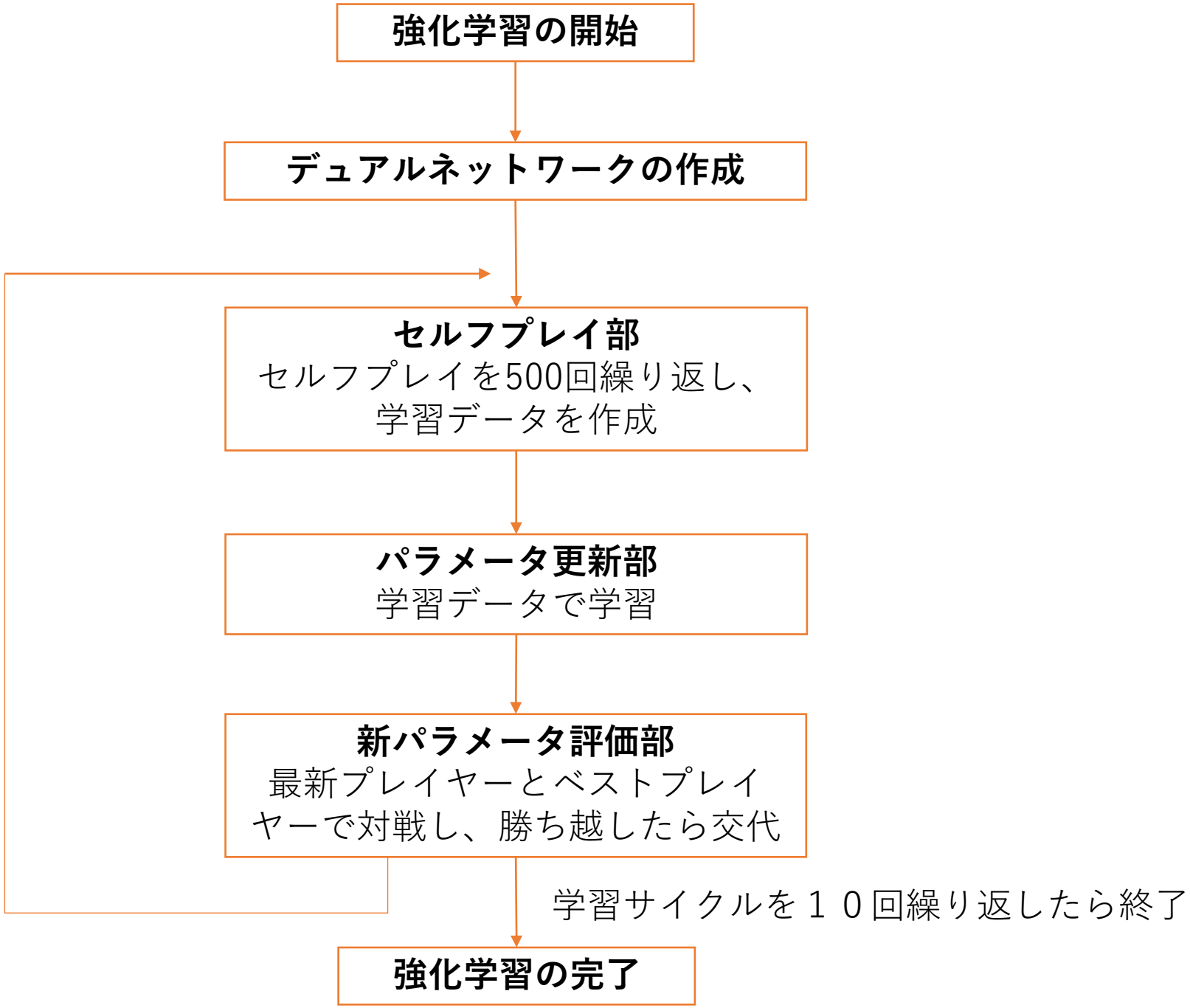
AlphaZeroのモンテカルロ木探索ー行動選択（学習時）



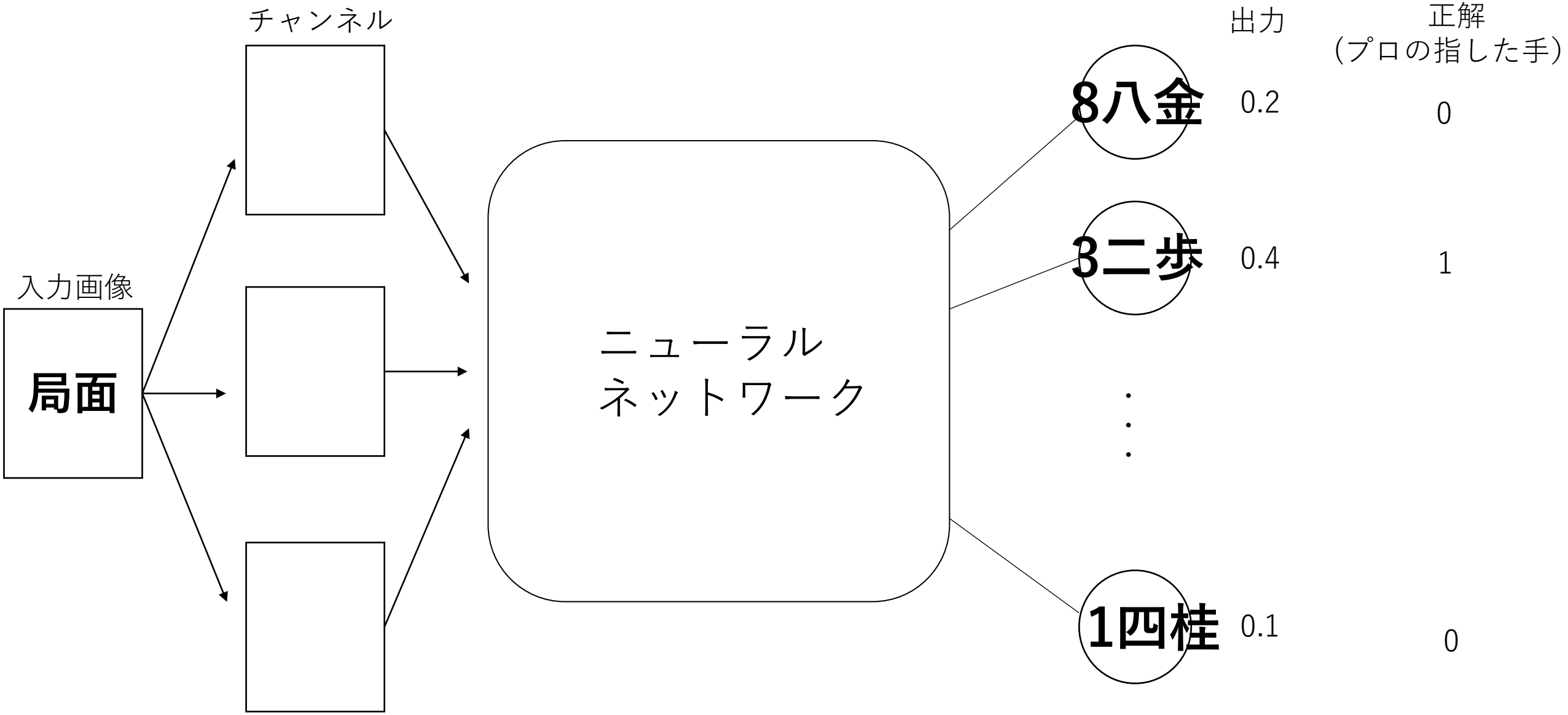


この確率分布に従って１手を選ぶ

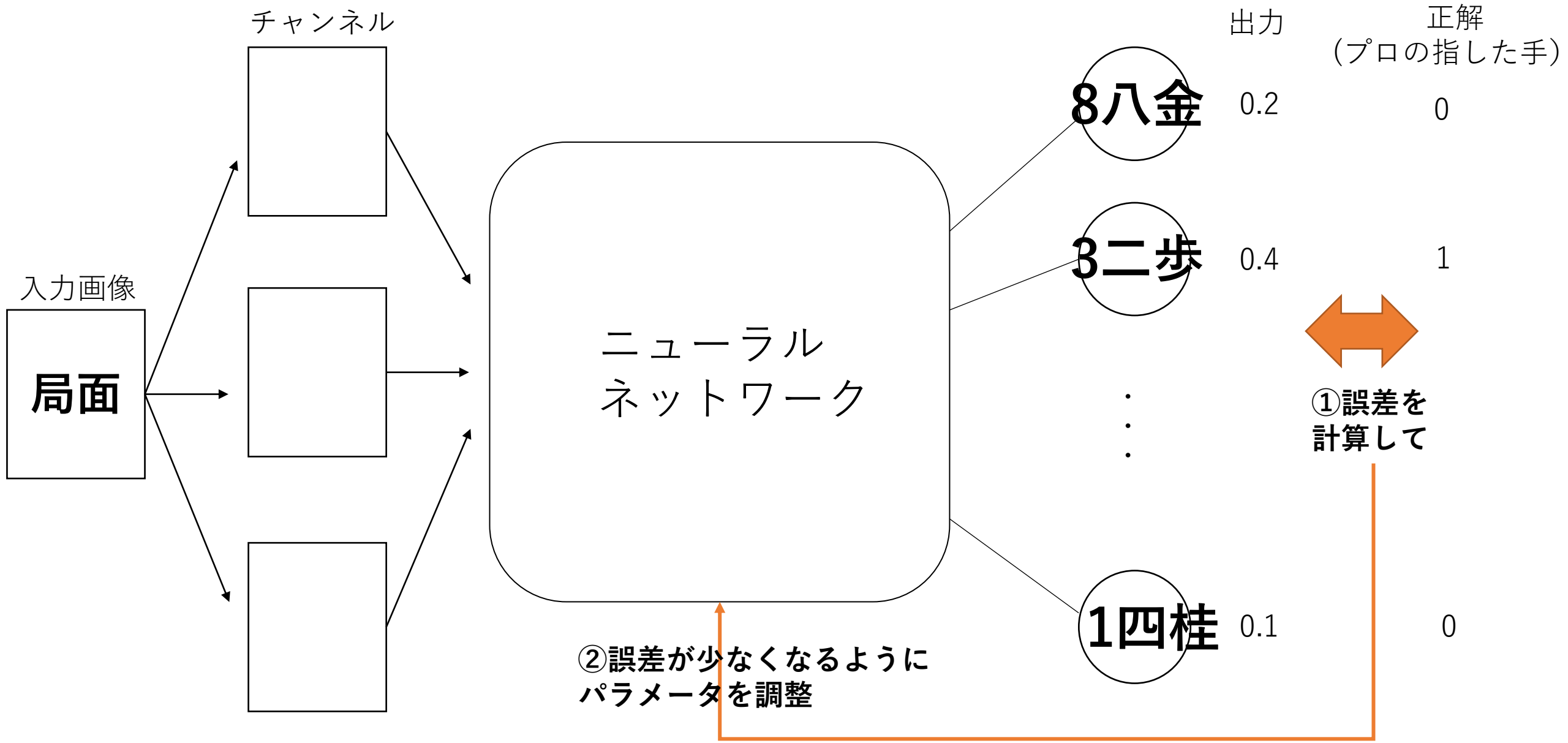
学習


強化学習の開始
デュアルネットワークの作成
セルフプレイ部
セルフプレイを500回繰り返し、
学習データを作成
パラメータ更新部
学習データで学習
新パラメータ評価部
最新プレイヤーとベストプレイ
ヤーで対戦し、勝ち越したら交代
学習サイクルを１０回繰り返したら終了
強化学習の完了

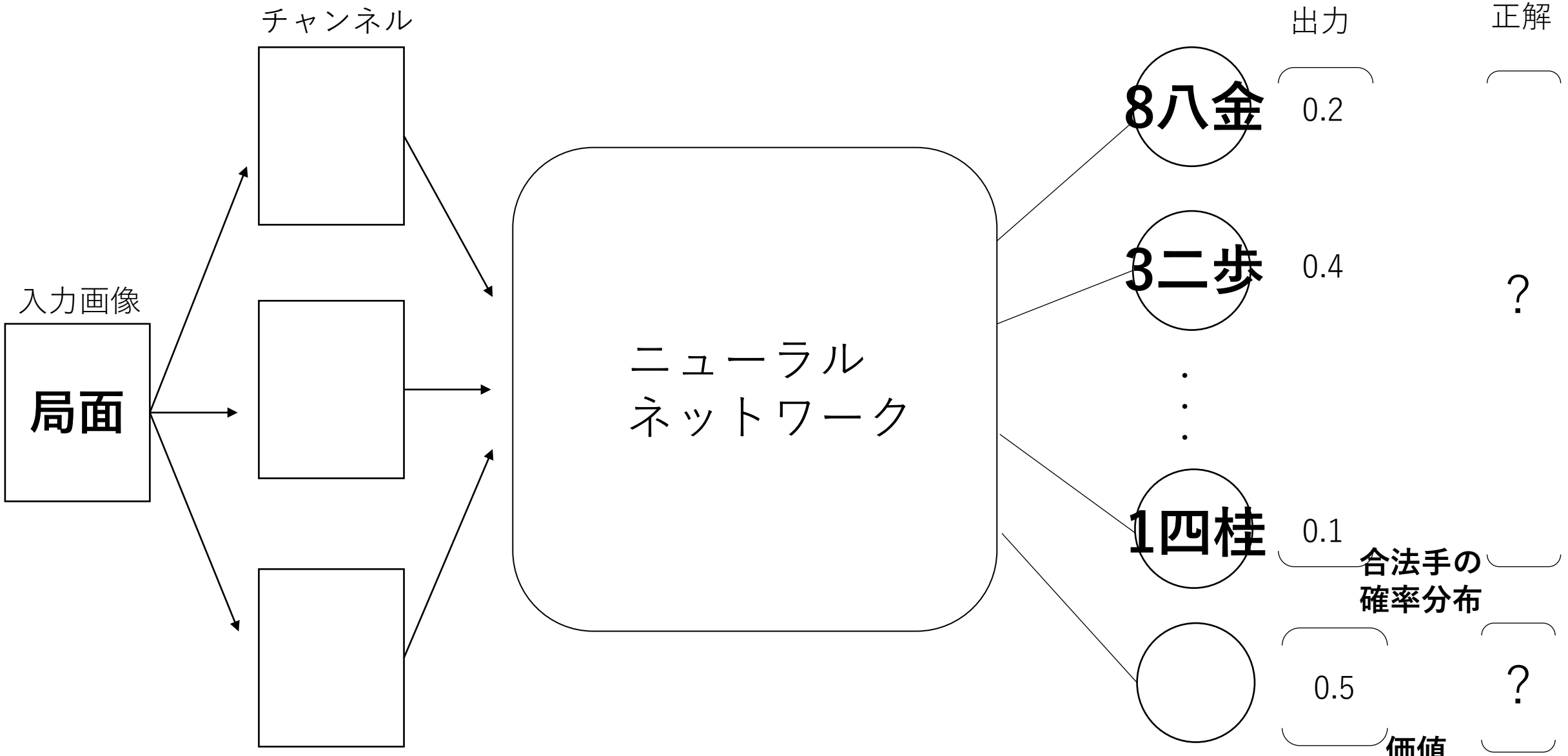
ニューラルネットワークの将棋への活用



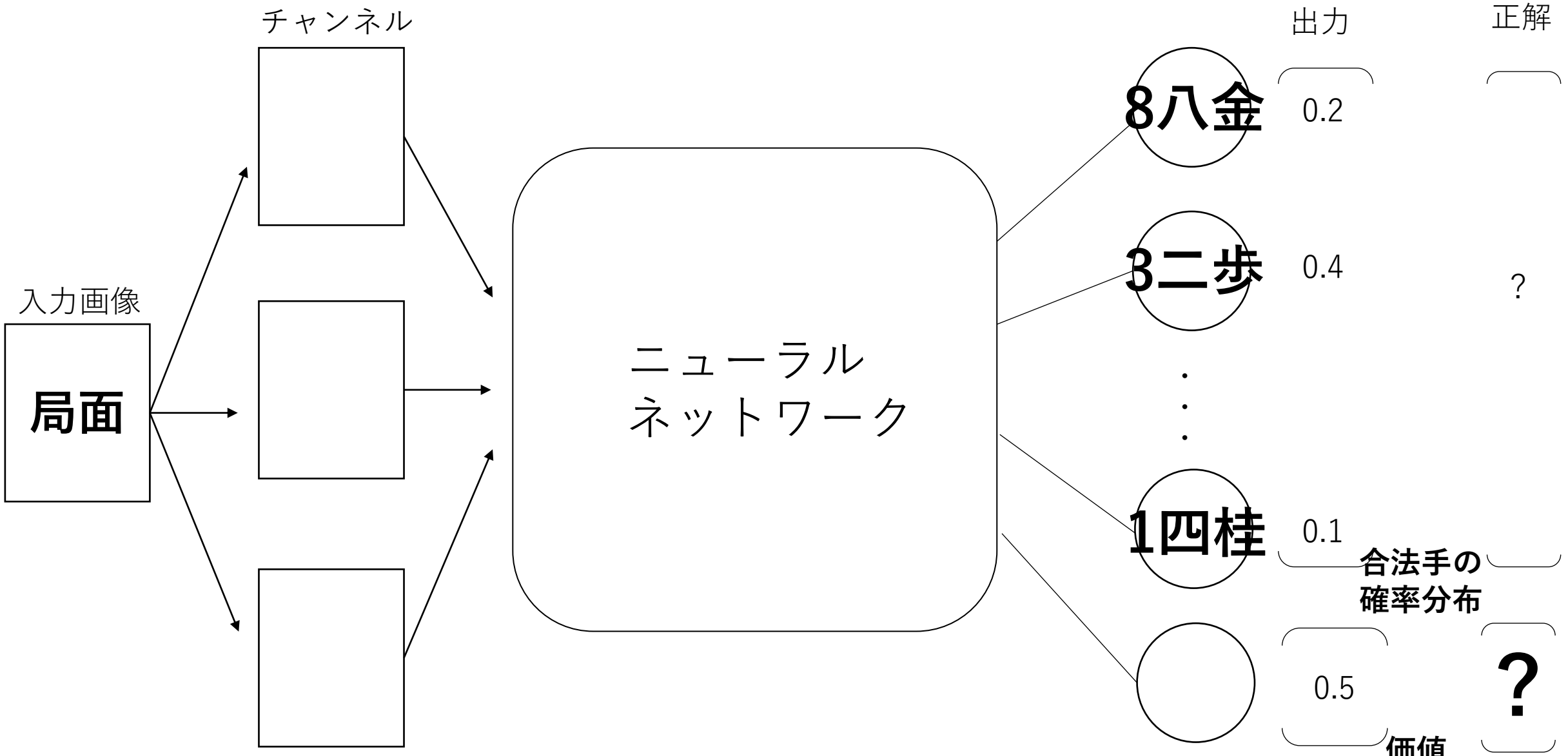
ニューラルネットワークの将棋への活用



AlphaZeroのニューラルネットワーク



AlphaZeroのニューラルネットワーク



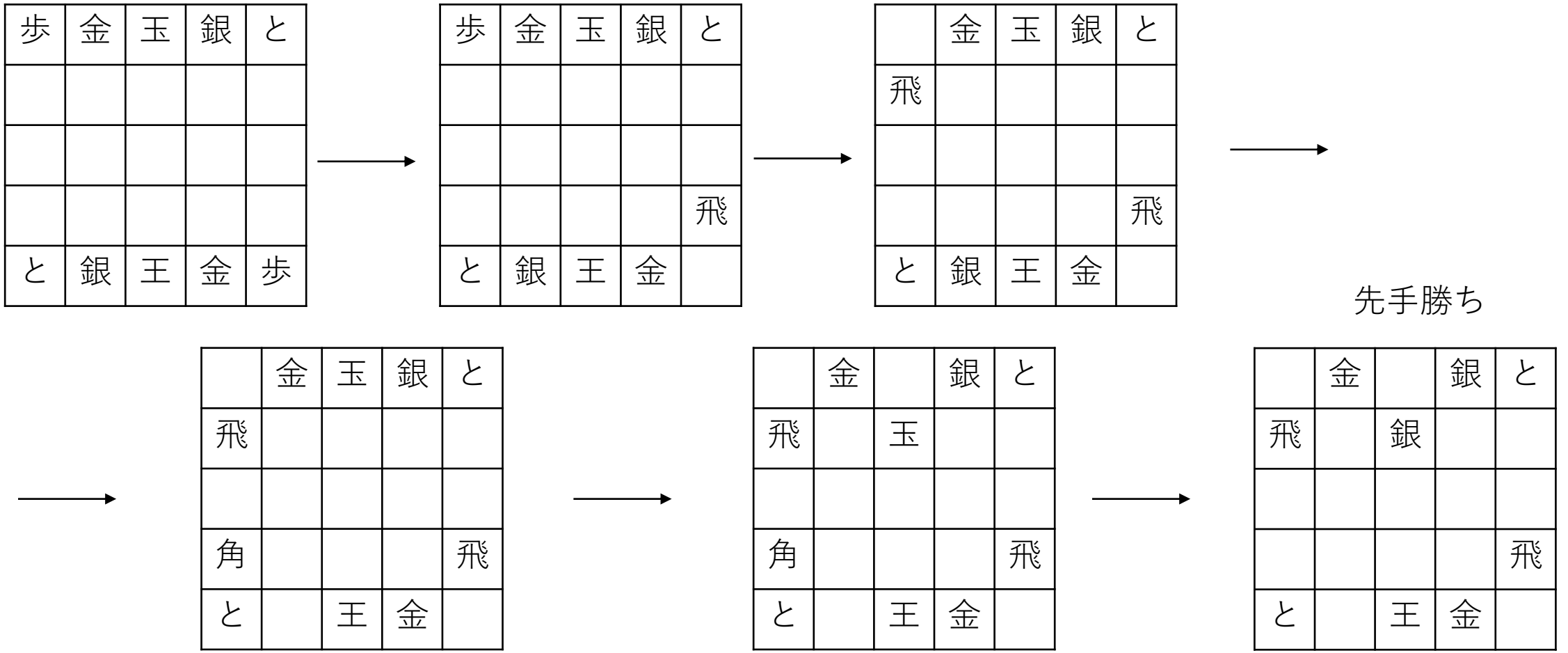
AlphaZeroのニューラルネットワークの「価値」の正解データ

勝った場合はその勝ちに至った局面の「価値」を1、
負けた場合は負けに至った局面を-1として学習させる。

学習後にさらに対戦させてデータをとって、
そのデータでさらに学習させてさらにさらに対戦させて
データを取って……ってのを繰り返してよいなら、こんな
に単純なやり方でも多くの場合でうまくいく（らしい）

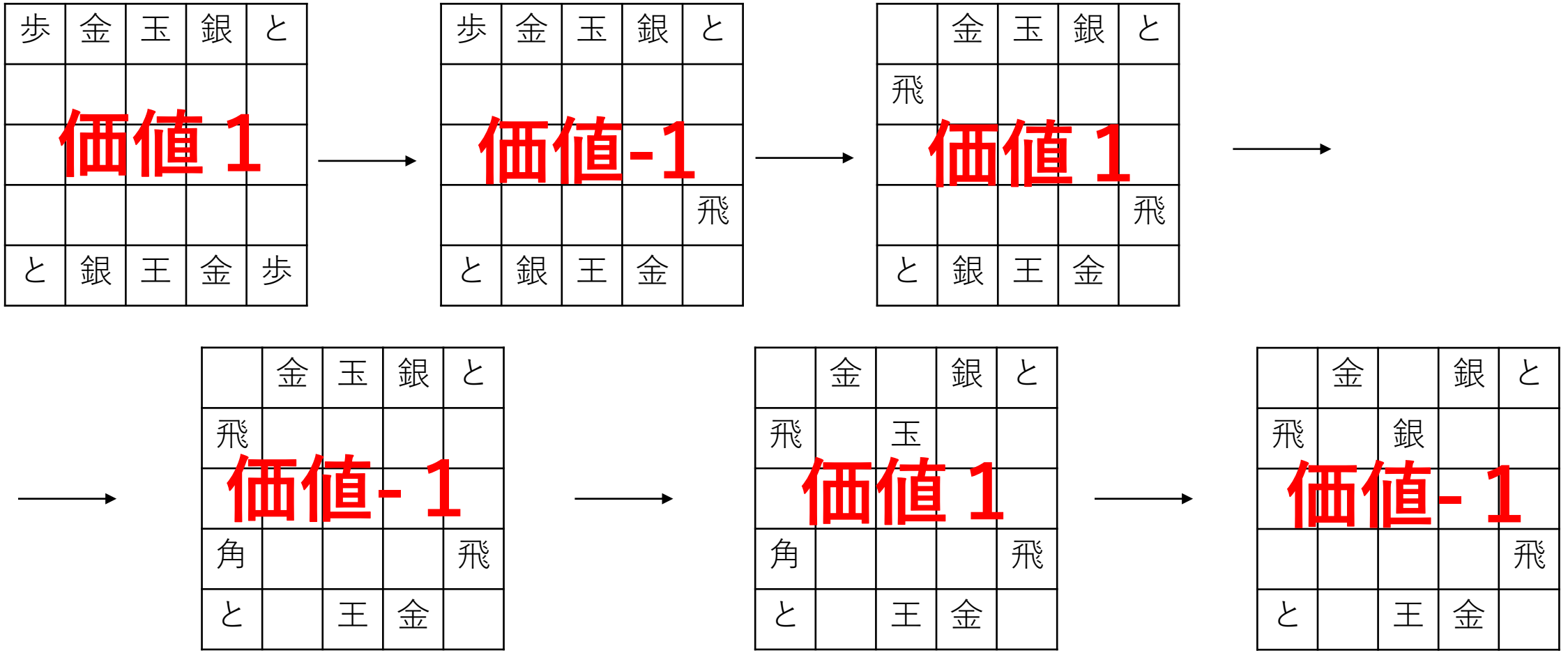
AlphaZeroのニューラルネットワークの「価値」の正解データ

勝った場合はその勝ちに至る局面の「価値」を1、
負けた場合は-1を正解とする。

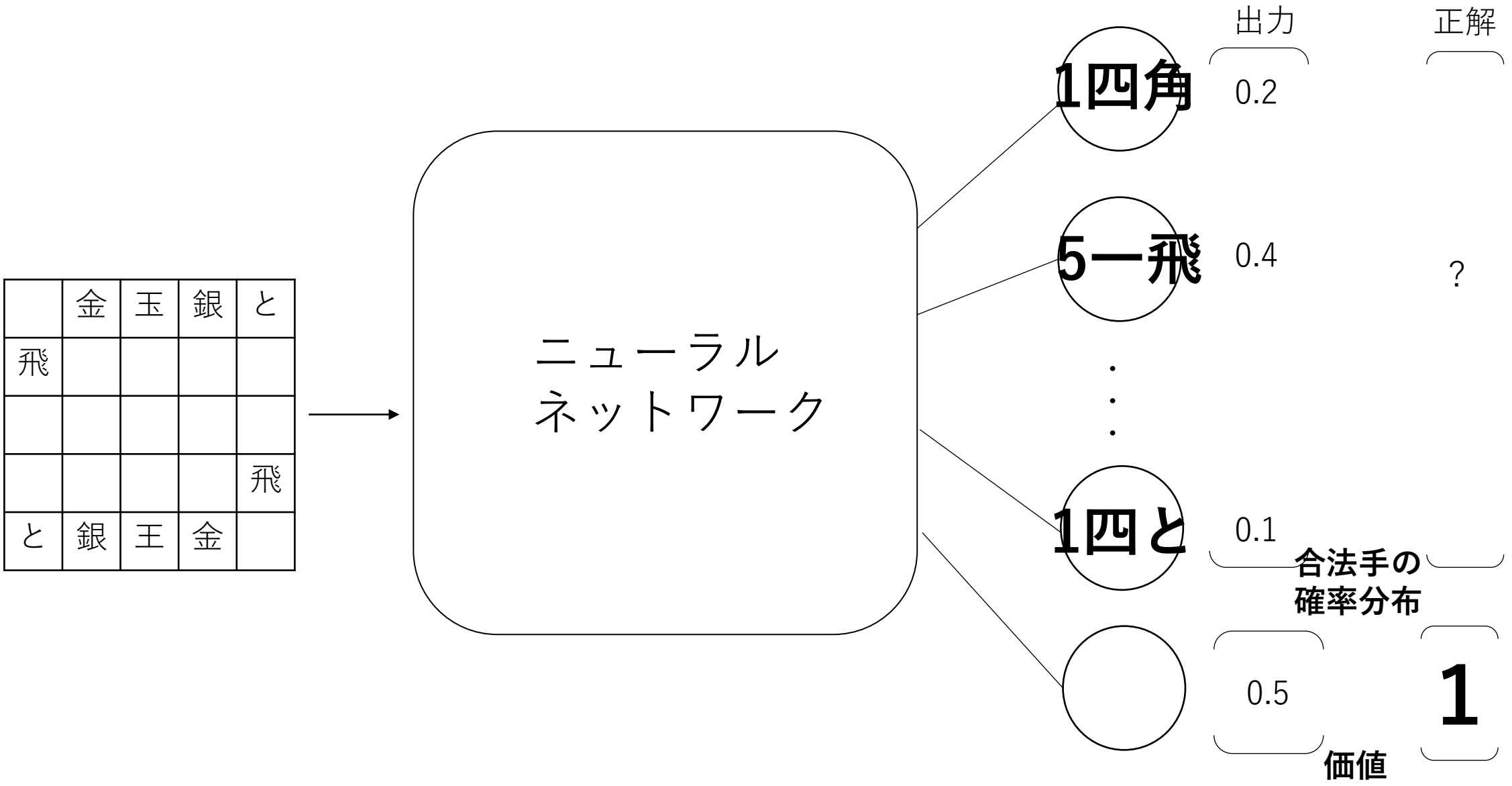


AlphaZeroのニューラルネットワークの「価値」の正解データ

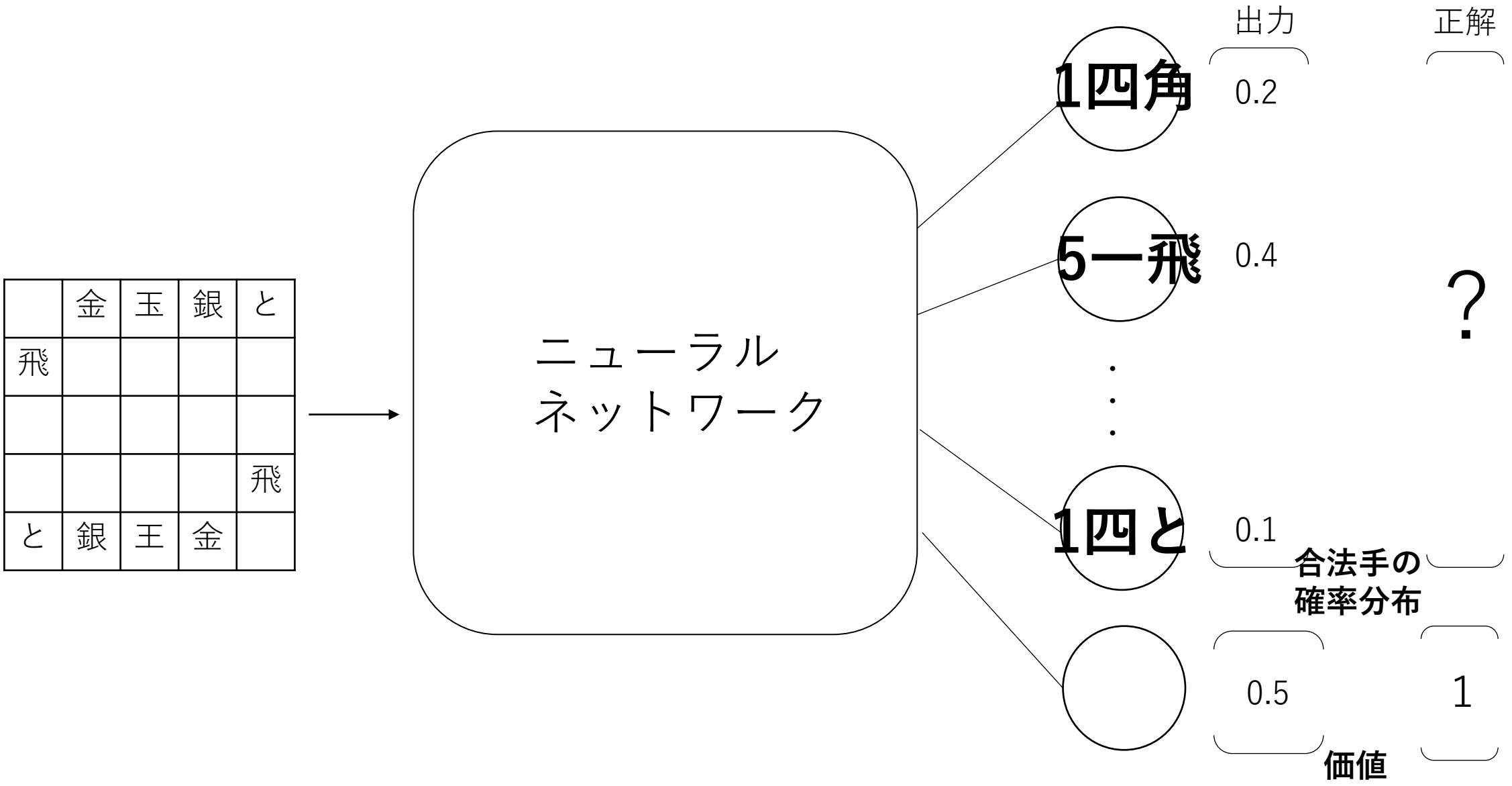
勝った場合はその勝ちに至る局面の「価値」を1、
負けた場合は-1を正解とする。



AlphaZeroのニューラルネットワーク



AlphaZeroのニューラルネットワーク



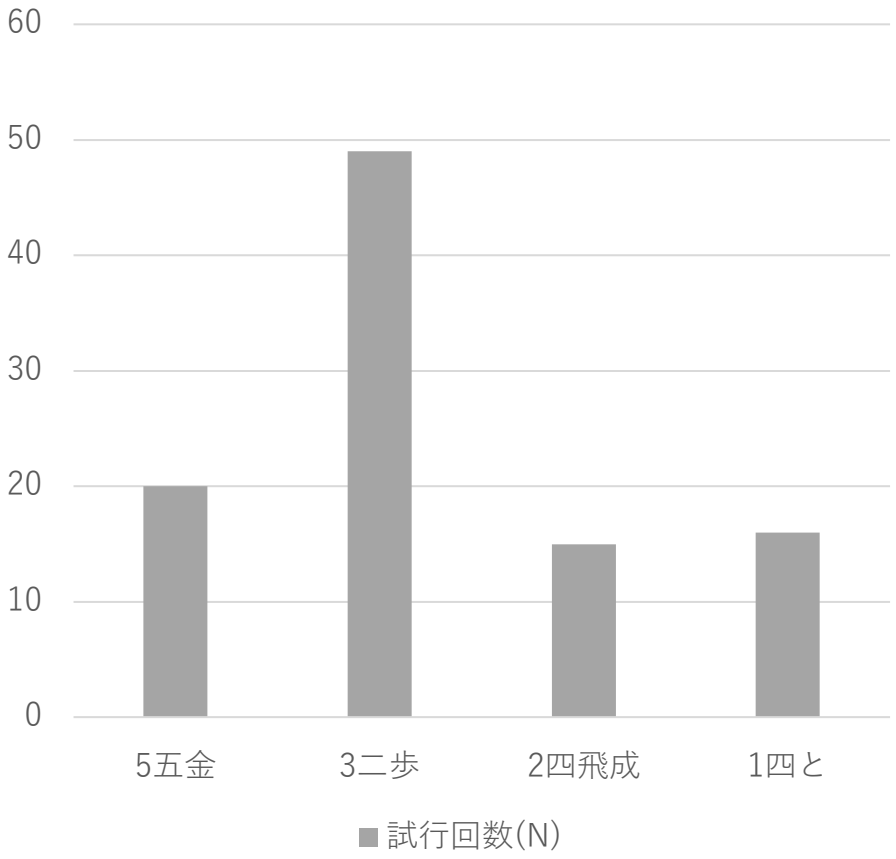
AlphaZeroのニューラルネットワークの「方策」の正解データ

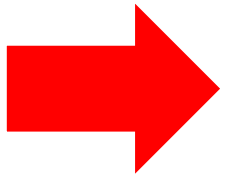
モンテカルロ木探索の試行回数の確率分布の値を正解にする。

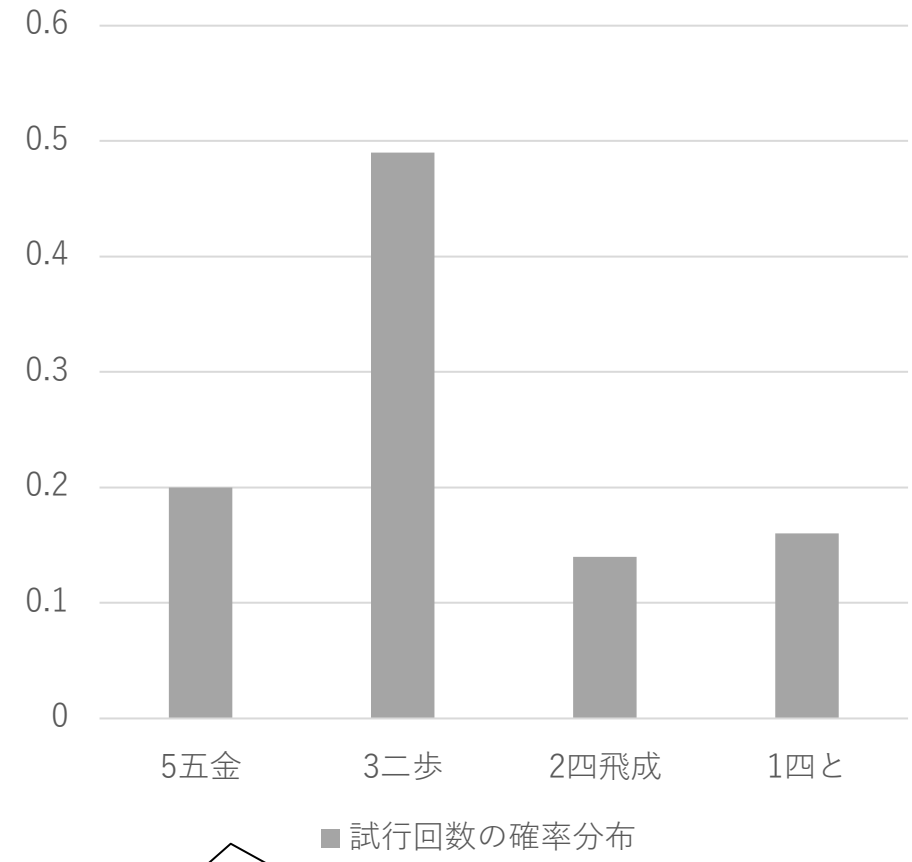
勝ったにせよ負けたにせよ試行回数の値はモンテカルロ木探索を行ったため、
深く読むべき局面と浅く読むべき局面を表現している、正解に近い値だと考える。

また、価値が正しく設定されるなら、その価値に沿う形に次回の対戦での
試行回数が変わって、次は今よりもさらに正解に近い値となる。

AlphaZeroのモンテカルロ木探索ー行動選択（学習時）（再掲）

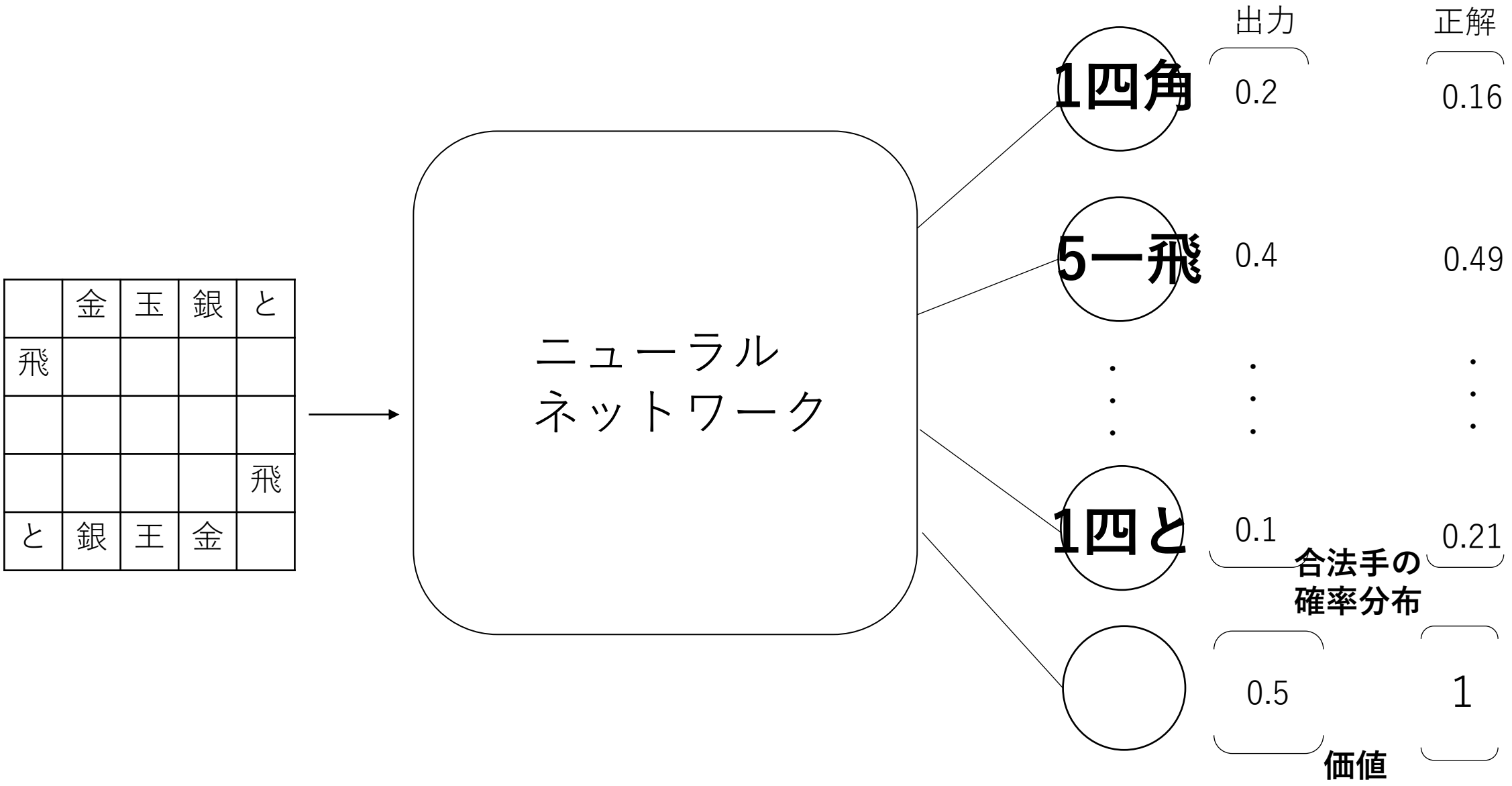






この確率分布に従って１手を選ぶ

AlphaZeroのニューラルネットワーク



目次

結果

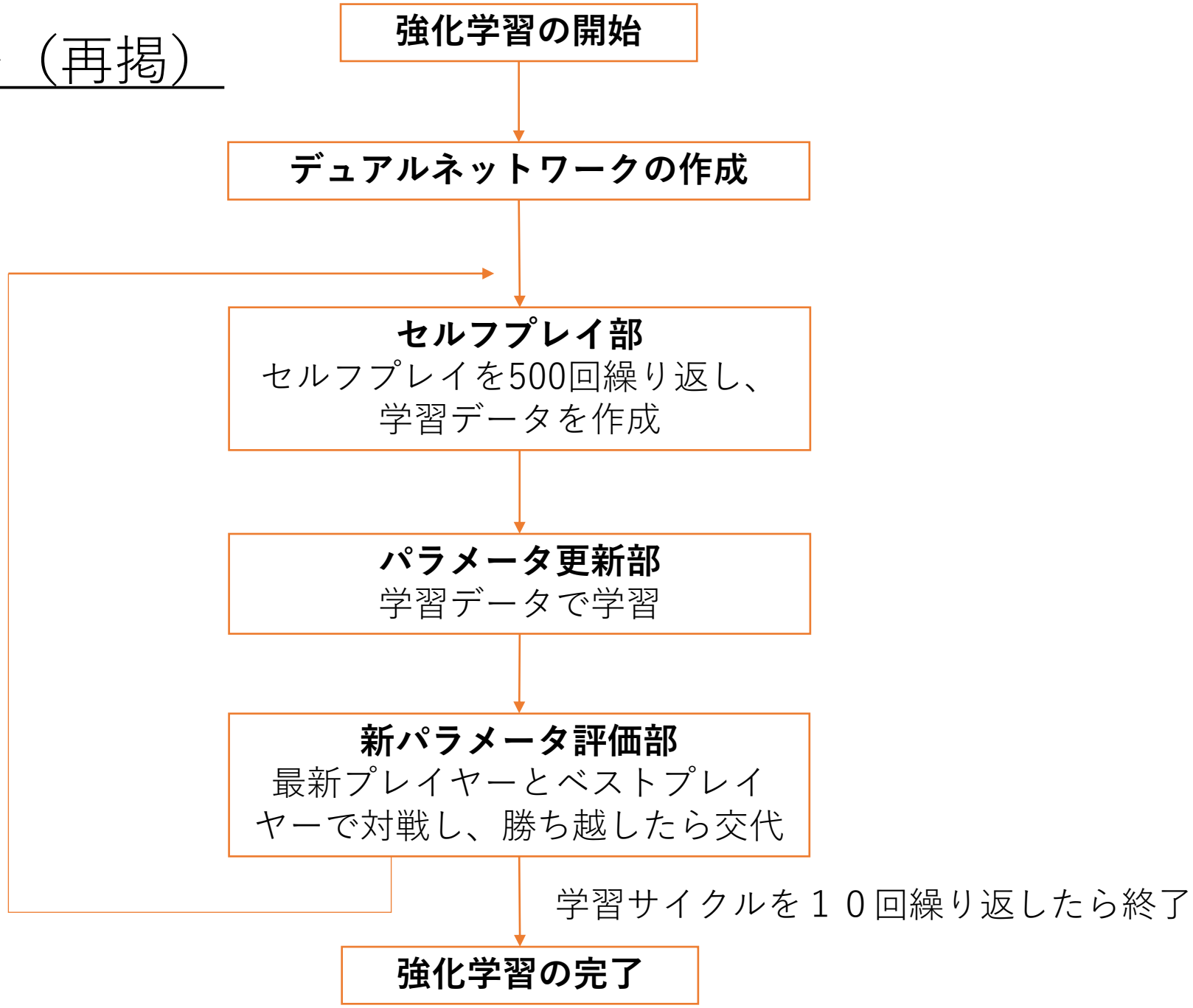
- １０サイクル
  VS モンテカルロ木探索 0.78
  VS AlphaBeta 0.20

- ２０サイクル
  VS モンテカルロ木探索 0.86
  VS AlphaBeta 0.22

- ３０サイクル
  VS モンテカルロ木探索 0.94
  VS AlphaBeta 0.21

- ４０サイクル
  VS モンテカルロ木探索 0.95
  VS AlphaBeta 0.18

学習サイクル（再掲）


強化学習の開始
デュアルネットワークの作成
セルフプレイ部
セルフプレイを500回繰り返し、
学習データを作成
パラメータ更新部
学習データで学習
新パラメータ評価部
最新プレイヤーとベストプレイ
ヤーで対戦し、勝ち越したら交代
学習サイクルを１０回繰り返したら終了
強化学習の完了

目次

参考文献

- 布留川栄一, 『AlphaZero 深層学習・強化学習・探索・人工知能プログラミング実践入門』, 株式会社ボーンデジタル, 2019.

- 曽我部東馬, 『強化学習アルゴリズム入門ー「平均」からはじめる基礎と応用ー』, オーム社, 2019.

- 『今年49歳になるおっさんでも作れたAlphaZero』, 閲覧日 2020.-1-30, https://tail-island.github.io/programming/2018/06/20/alpha-zero.html