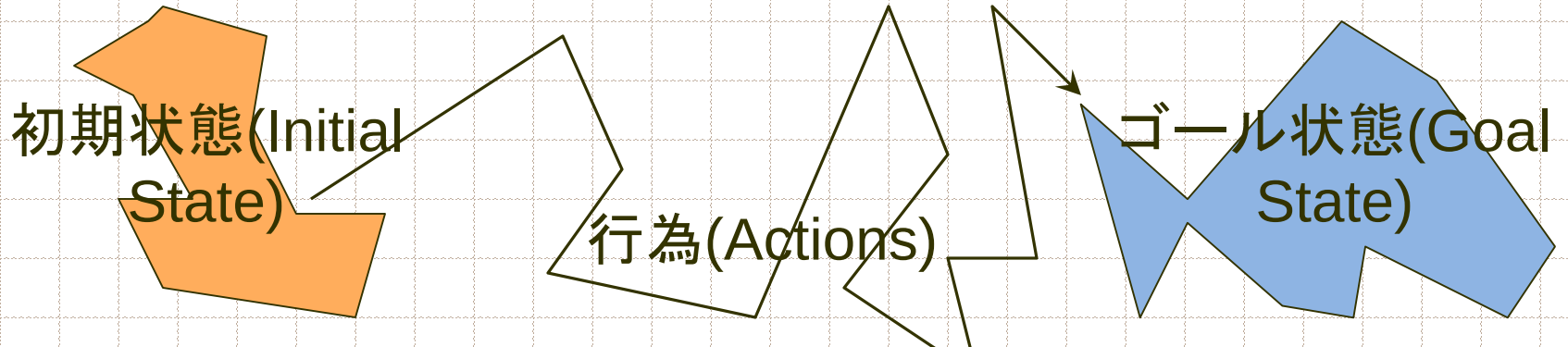
人工知能基礎
第14回

探索による問題解決(6)、
モンテカルロ木探索

ソフトウェア情報学部
David Ramamonjisoa

ゴールを用いるエージェントの構築

- エージェントの環境を状態で表現する(**state**)?
- エージェントのゴールは何か(**goal** to be achieved?)
- 可能な行為は何かある(What are the **actions**?)
- 問題を解決するためにどのような情報が状態と状態遷移に記述するべきか



探索戦略

- 探索戦略選択のための四つの基準
  - 完全性： 解が存在するとき，それを見つけることが保証されているか？
  - 最適性： いくつか異なる解があるとき，戦略は最も良い解を見つけるか？
  - 時間計算量： 解を見つけるまでにどれくらい時間が掛かるか？
  - 空間計算量： 探索を行うためにどのくらいメモリを必要とするか？
- 情報を持つ探索(informed search)，ある探索
  - 現在の状態からゴールに至る順路の中で、最小コストの順路（最適順路）などを考えて効率的に行う.

ゲームプレイング

- 将棋、チェス、チェッカー、囲碁、オセロのような、二人ゲームを考える （ボードゲーム）
- そのようなゲームのコンピュータプログラムは、どうつくれば良いのか。
  - 探索問題の一種。
  - 良さそうな手を打つ。明らかに無駄な手は打たない。
  - 相手がこちらを妨害する。そのような状況で、最善の手を考える。
- どれくらい強いプログラムが作られたか。

情報をもつ探索戦略(Informed search strategies)

- 情報をもつ探索戦略は問題定式化の使用可能情報のみ利用する。
- 敵対的探索(adversarial search)
  - 相手の手番が予測不可能ため、可能な手番をすべて探索する
  - ゲームの時間制限があるため、手番(ゴール)を近似する

目次

- ゲームプレイング
  - 原始モンテカルロ探索
  - モンテカルロ木探索

原始モンテカルロ探索

- アルファベタ法で探索木を枝刈りしたとはいえ、三目並べは再帰的にゲーム終了まで調べて計算していた。オセロゲームでは深さ10を制限し、探索を行った。
- 局面が多いゲームは膨大な時間がかかり、終了までのゲーム木探索は現実的ではありません
- 新しい手法：手作りの評価関数、原始モンテカルロ探索

原始モンテカルロ探索

- ランダムシミュレーションによって状態価値を計算する方法
- 現在の局面からゲーム終了まで何回もランダムプレイを行い、その勝率の高い手を価値が高いと見なす
- アルゴリズム：①局面を作成、②ランダムで状態価値計算、③アルファベタ法で状態価値を計算、④プレイアウト、⑤原始モンテカルロ探索で行動選択
- 評価

①局面を作成

- ミニマックス法と同様に行う
- ゲーム木を生成する
- 深さ優先探索でリーフノードから、ルールに従って状態評価を計算
  - 先手局面のノードは、その子ノードの状態評価の最大値を状態価値とする
  - 後手局面のノードは、その子ノードの状態評価の最小値を状態価値とする

②ランダムで状態価値計算

- ゲーム木を生成する
- 深さ優先探索でリーフノードから、ルールに従って状態評価を計算
  - 先手局面のノードは、合法手の中でランダムで状態を選択する
  - 後手局面のノードは、合法手の中でランダムで状態を選択する
  - ゲーム終了時、状態価値「-1：負け」、「0：引き分け」を返す

③アルファベタ法で状態価値を計算

- 状態を渡すと評価値を返す
- ゲームの途中はベストスコア（アルファベタ法）を返す
- 深さ優先探索でリーフノードから、ルールに従って状態評価を計算
  - ゲーム終了時、状態価値「-1：負け」、「0：引き分け」を返す

④プレイアウト

- 現在の局面からゲーム終了までプレイすることを「プレイアウト」と呼ぶ
- 深さ優先探索でリーフノードから、ルールに従って状態評価を計算
  - ゲーム終了時、状態価値「1：勝ち」、「-1：負け」、「0：引き分け」を返す
- チェスでは1回につき80手が1つのプレイアウトになる

⑤原始モンテカルロ探索で行動選択

- 合法手ごとに、10回プレイアウトした時の状態価値の合計を計算する。そして、合計が最も大きな行動を選択する
- プレイアウトの回数は多いほど精度は増する、変わりに時間がかかる

評価

- 原始モンテカルロ探索　vs　アルファベタ法
  - 100回ゲームをプレイしてその勝率を計算
  - 先手と後手を交互に交代する
- 原始モンテカルロ探索はアルファベタ法に負けていることある。
- 割合ではアルファベタ法が勝利である

モンテカルロ木探索

- 原始モンテカルロ法をさらに改善した探索手法
- モンテカルロ木探索は強化学習を使われる手法
- シミュレーション
- アルゴリズム：①選択、②評価、③展開、④更新

モンテカルロ木探索

- ランダムシミュレーションによって状態価値を計算する方法
- 「モンテカルロ木探索」の「ゲーム木」の「初期状態」は「ルートノード」（現在の局面）とその「子ノード」のみから始まる

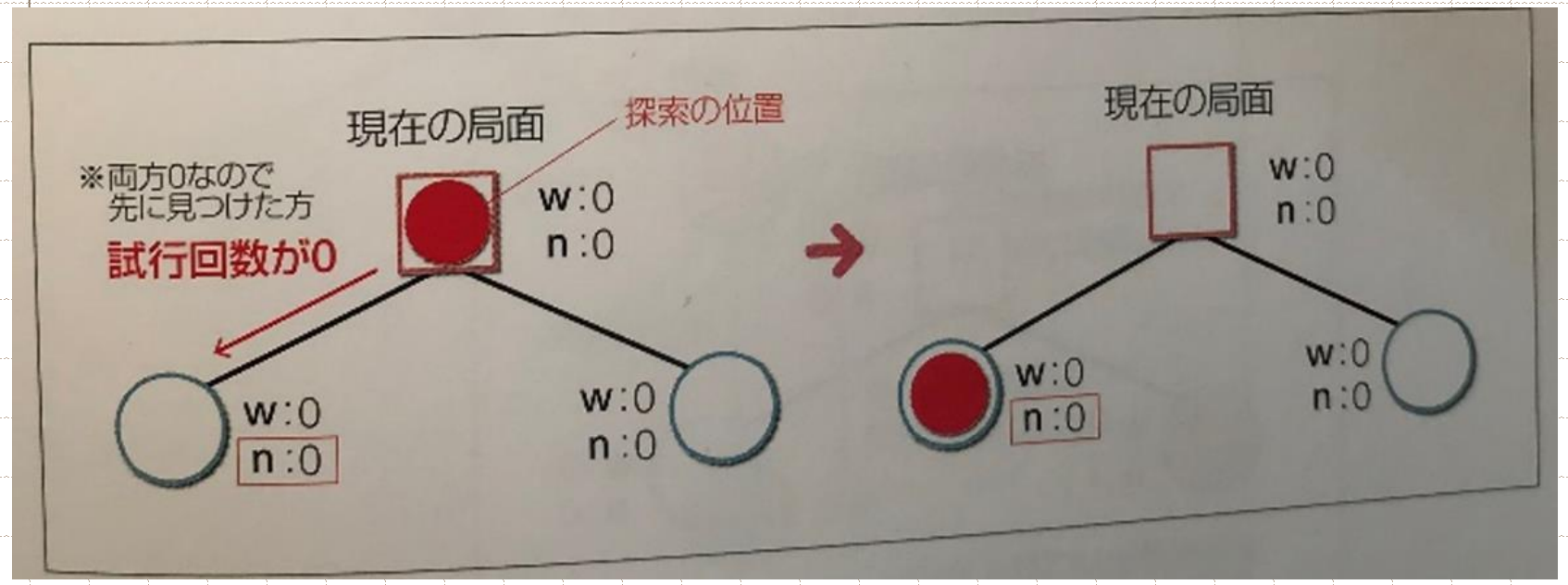
①選択

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

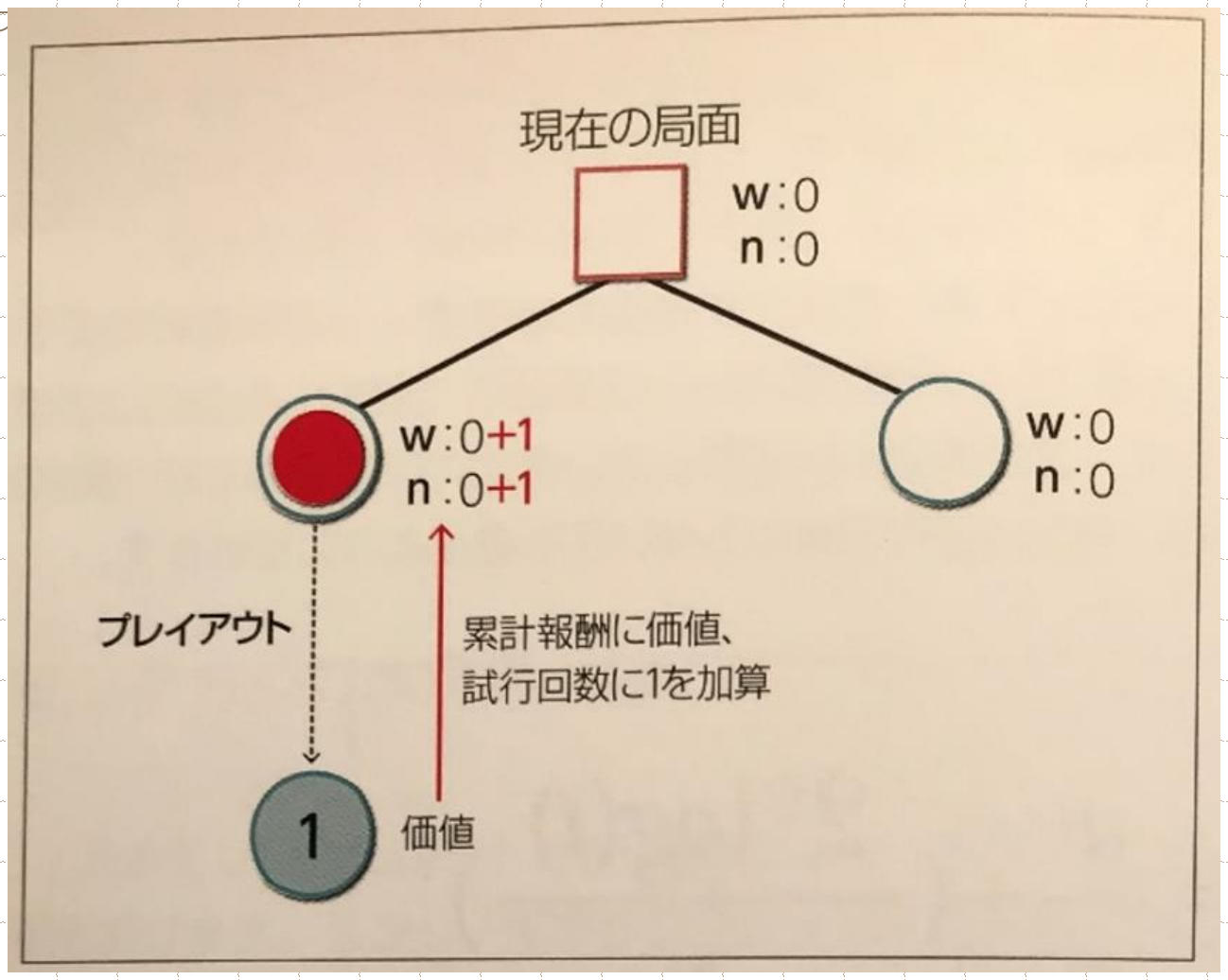
成功率　　バイアス

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の合計

- UCB1:Upper Confidence Bound 1



②評価



③展開

展開

プレイアウト後に、「リーフノード」の試行回数が任意の回数以上（今回は10回とします）となったら、そのノードが持つ合法手を子ノードとして追加します。この操作を「展開」（Expansion）と呼びます。

初回のシミュレーションでは、「リーフノード」の試行回数はまだ「1」なので「展開」しません。

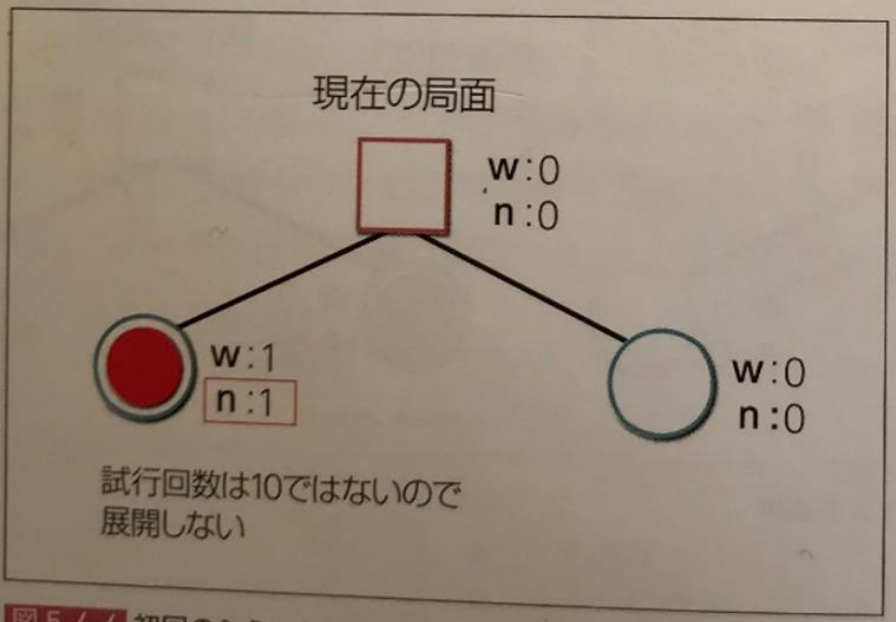


図 5-4-4 初回のシミュレーションの展開

④更新

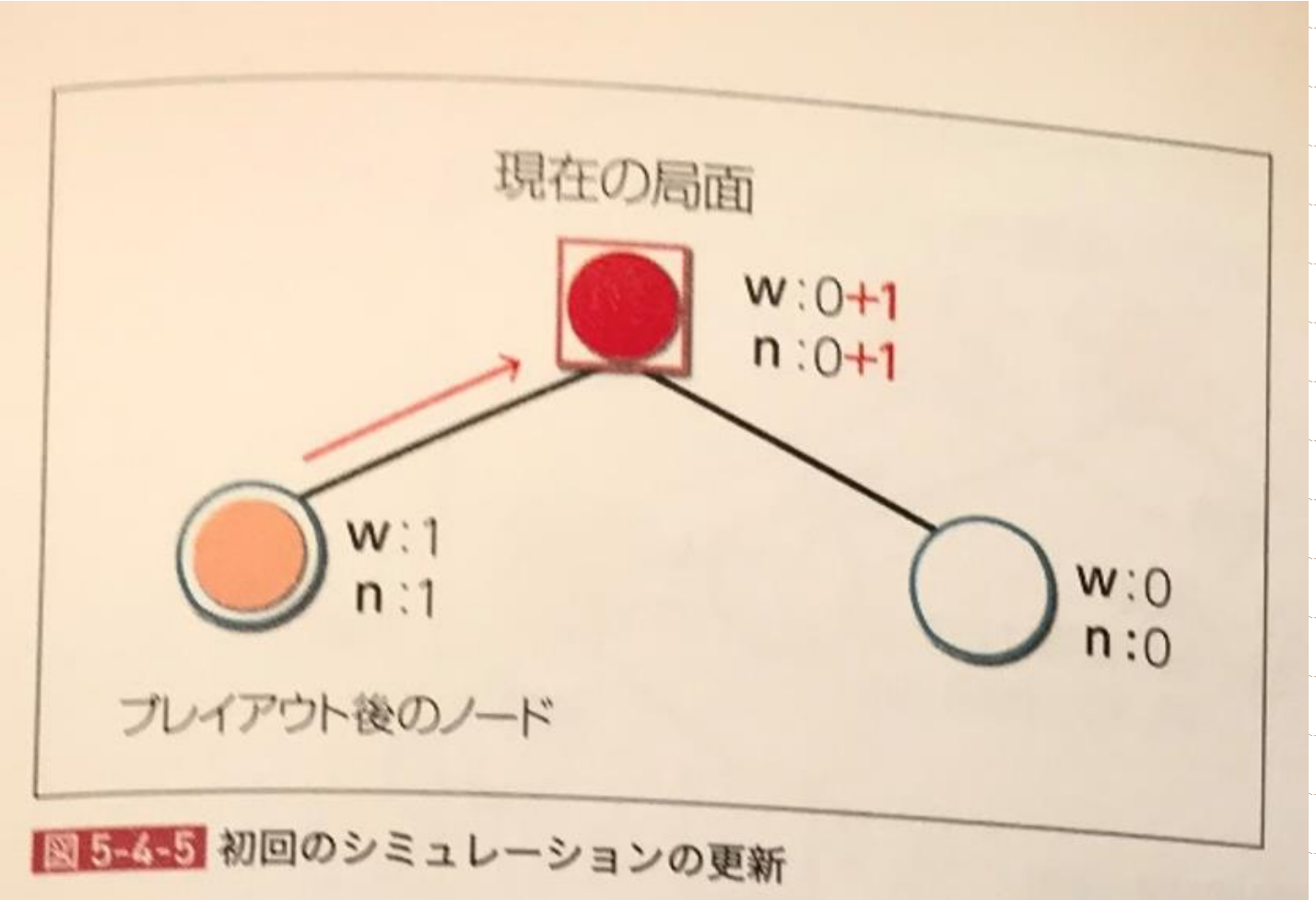


図 5-4-5 初回のシミュレーションの更新

シミュレーション2

2回目のシミュレーション

「ルートノード」から探索を開始し、「選択」「評価」「展開」「更新」の4つの操作で探索を行い、「ルートノード」まで戻ってくることで、シミュレーション1回分となります。

2回目のシミュレーションでは、「選択」で試行回数0の右のリーフノードを選択します。そして、「プレイアウト」を実行後、「展開」は試行回数が10ではないので展開せず、「更新」で「累計価値」と「試行回数」を更新します。

この例では「負け」なので、価値「－1」で更新しています。

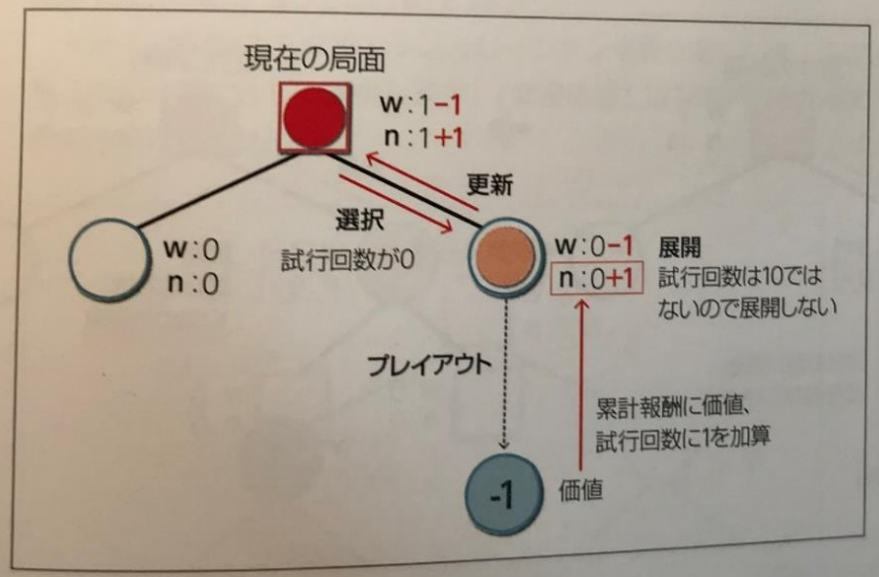
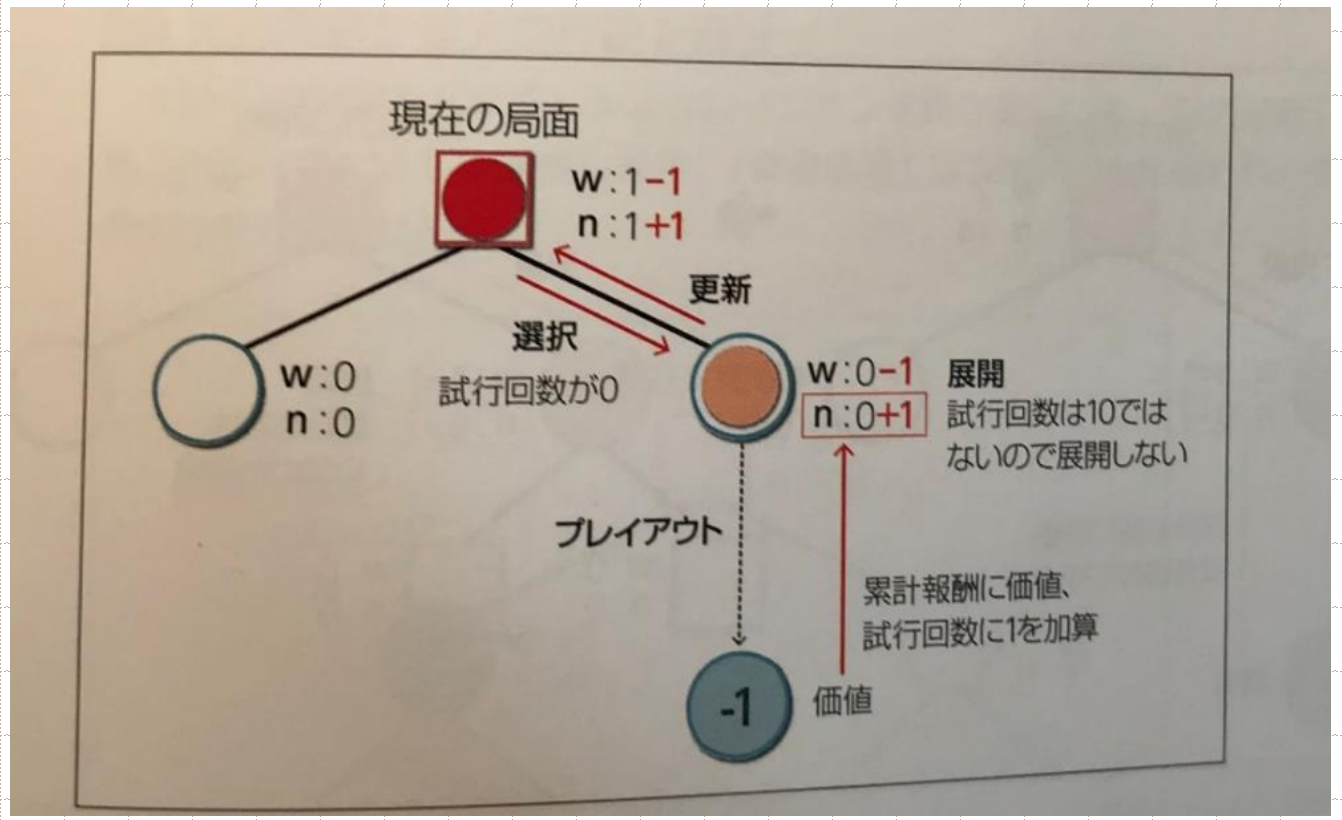
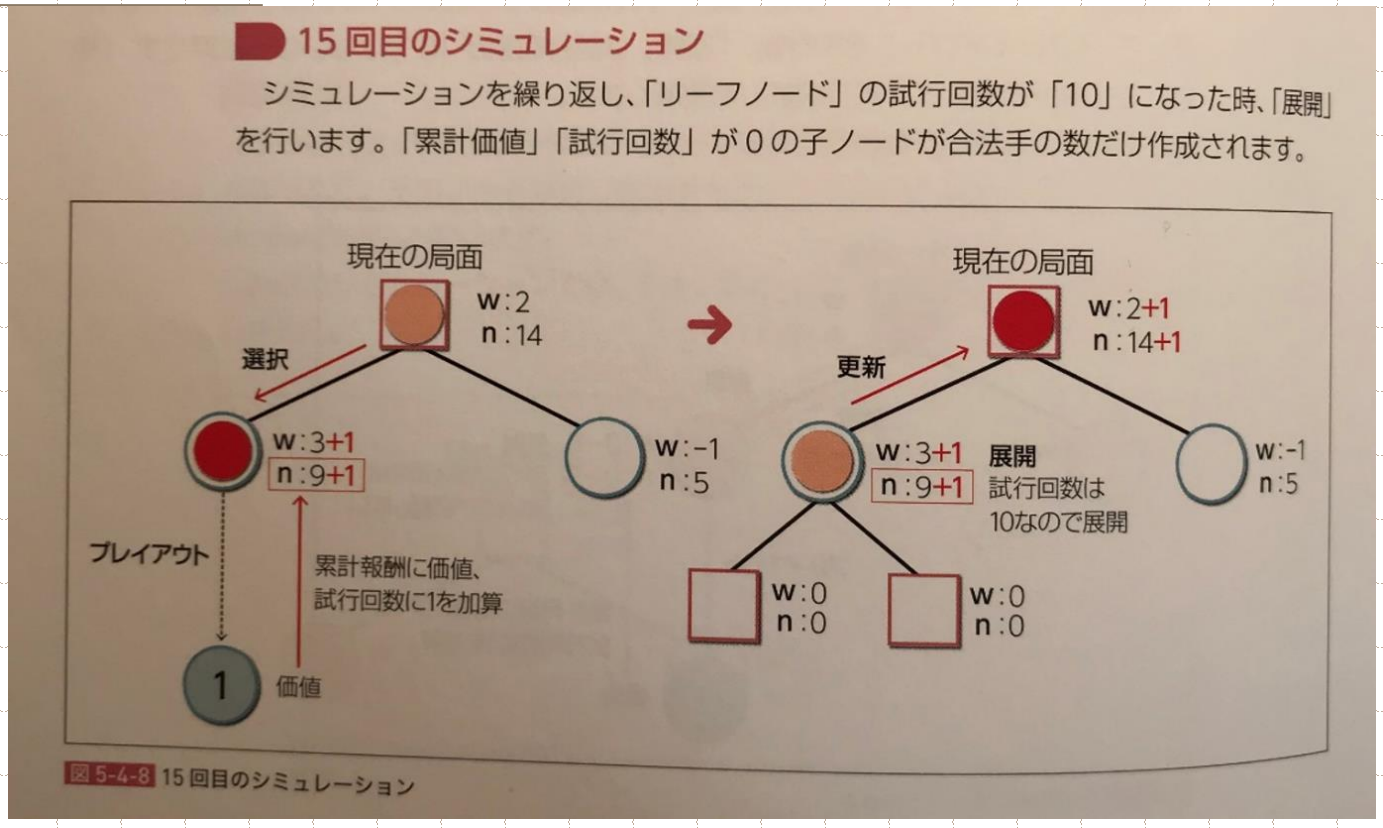


図 5-4-6 2回目のシミュレーション

シミュレーション3


現在の局面
w:1−1
n:1+1
更新
選択
w:0
n:0
試行回数が0
w:0−1
n:0+1
展開
試行回数は10では
ないので展開しない
プレイアウト
累計報酬に価値、
試行回数に1を加算
-1
価値

シミュレーション15


15回目のシミュレーション
シミュレーションを繰り返し、「リーフノード」の試行回数が「10」になった時、「展開」を行います。「累計価値」「試行回数」が0の子ノードが合法手の数だけ作成されます。
現在の局面
w:2
n:14
選択
w:3+1
n:9+1
w:-1
n:5
プレイアウト
累計報酬に価値、
試行回数に1を加算
1
価値
現在の局面
w:2+1
n:14+1
更新
w:3+1
n:9+1
展開
試行回数は
10なので展開
w:-1
n:5
w:0
n:0
w:0
n:0
図5-4-8 15回目のシミュレーション

シミュレーション16

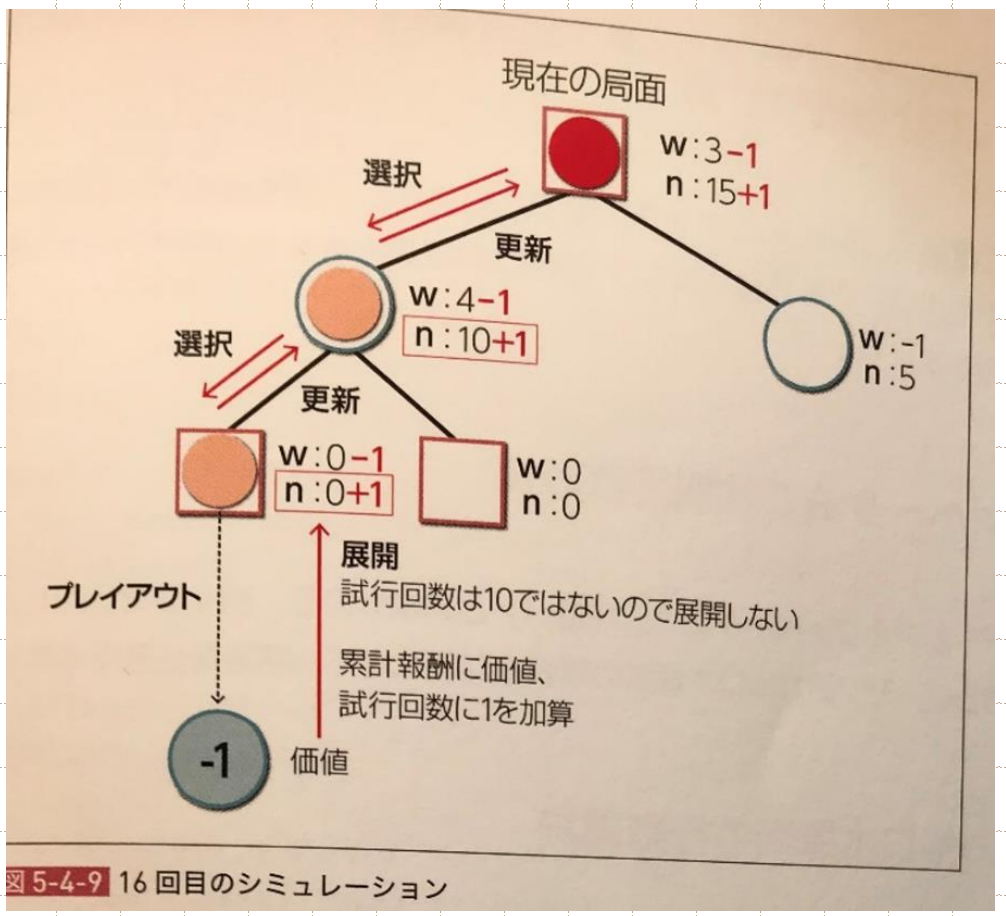


図 5-4-9 16 回目のシミュレーション

シミュレーション100

試行回数が最大の行動を選択

十分に（今回は 100 回）シミュレーションを繰り返した後、「試行回数」が最大の行動を「次の一手」として選択します。「累積価値」は探索時のみに使われ、最終的な行動選択には使われません。

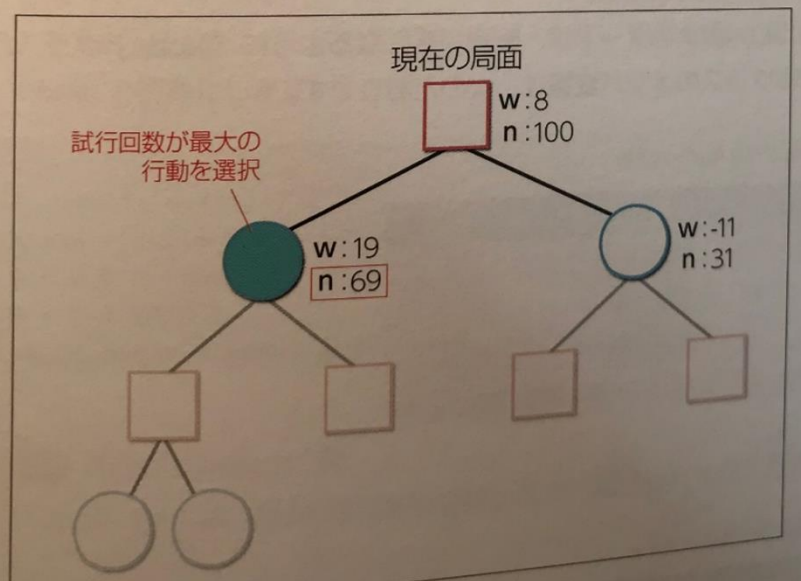


図 5-4-10 試行回数が最大の行動を「次の一手」として選択

評価

- モンテカルロ木探索　vs　アルファベタ法
  - 100回ゲームをプレイしてその勝率を計算
  - 先手と後手を交互に交代する
- モンテカルロ木探索はアルファベタ法にたまに負けている。
- 割合ではモンテカルロ木探索が勝利である

AlphaGo Zero（アルファゼロ）

- DeepMindの囲碁ソフトウェア
- 2017年10月19日に学術誌Natureの論文でAlphaGo Zeroを発表した
- 人間の対局からのデータを使わずに作られており、それ以前の全てのバージョンよりも強い

AlphaZero（アルファゼロ）

- DeepMindによって開発されたコンピュータプログラムである
- 汎化されたAlphaGo Zeroのアプローチ
- AlphaZeroは24時間以内にチェス、将棋、囲碁の世界チャンピオンプログラムであるStockfish、elmo、3日間学習させたAlphaGo Zeroを破るレベルに達した

ゲーム探索のまとめ

- 相手がある場合には、自分、敵とも、最強の手を打とうとする。
- 自分の評価値を基準にすれば、自分は評価値を最も高くし、相手は、それを最も低くするように振舞う。
- 原始モンテカルロ法について解説した。
- アルゴリズムにα-β法と比較。
- モンテカルロ木探索の説明、評価
- 先端技術アルファゼロの概要

参考ページ,資料

- https://ja.wikipedia.org/wiki/ゲーム木
- https://ja.wikipedia.org/wiki/ミニマックス法
- https://ja.wikipedia.org/wiki/アルファ・ベータ法
- https://ja.wikipedia.org/wiki/**モンテカルロ木探索**
- https://ja.wikipedia.org/wiki/AlphaGo_Zero
- https://ja.wikipedia.org/wiki/AlphaZero
- 最強囲碁AI アルファ碁 解体新書 増補改訂版 アルファ碁ゼロ対応 深層学習、モンテカルロ木探索、強化学習から見たその仕組み (AI & TECHNOLOGY)　大槻知史　出版社：翔泳社; 増補改訂版 (2018/7/17)
- 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布留川英一 (著), 佐藤 英一 (編集),出版社 :ボーンデジタル (2019/6/28)
- 日経ソフトウエア2019年11月号　AIと対戦するリバーシを作る（特集3）