

確率的ゲーム 2048 の強化学習の研究

山下修平

2023 年 12 月 8 日

目次

第 1 章	はじめに	3
第 2 章	2048	4
2.1	2048 のルールと用語説明	4
2.2	ゲームの進行上の性質	4
第 3 章	2048 と強化学習	7
3.1	強化学習の概要	7
3.2	AlphaZero	10
3.3	2048 への強化学習の応用	12
第 4 章	2048 と完全解析	13
4.1	2048 の完全解析とは	13
4.2	盤面サイズが小さな 2048 の完全解析	14
第 5 章	提案手法	17
5.1	3×3 盤面の 2048 の完全解析と強化学習	17
付録 A	実装の詳細	18
A.1	ゲーム環境の実装	18
A.2	完全解析の実装	18
A.3	強化学習の実装	18
参考文献		19

第 1 章

はじめに

第 2 章

2048

2.1 2048 のルールと用語説明

2048 は、Gabriele Cirulli によって公開された 1 人用のパズルゲームである [1]。ゲームは 16 マスからランダムに選ばれた 2 マスに 2 か 4 の数字タイルが置かれた盤面から始まる。プレイヤーが行うことは上下左右いずれかの方向を選択することである。プレイヤーがある 1 つの方向を選ぶと、盤面上のすべての数字タイルは選択した方向に向かってスライドして移動する。スライドする数字タイルは空きマスを通り、異なる数字タイルの直前か盤面の端で停止する。スライドして移動する際に 2 つの同じ数字のタイルが衝突すると、これらは合体してその合計の数字の 1 つのタイルへ変化し、プレイヤーはその数値を得点として獲得する。そのため、ゲームには 2 の累乗の数字タイルしか現れない。図 2.1 にある盤面から上下左右を選択したときの、数字タイルのスライドの仕方の具体例を示す。

数字タイルのスライド後、空きマスから等確率に選択されたある 1 マスに 90% の確率で 2 のタイルが、10% の確率で 4 のタイルが置かれる。ゲームはプレイヤーの行動による数字タイルのスライドと新たな数字タイルの出現を交互に繰り返して進行する。盤面上の数字タイルが市松模様になると、プレイヤーが選択可能な行動がなくなったときにゲームは終了する (図 2.2 を参照)。

ここでプレイヤーが行動を選択する盤面を**状態**、行動を選択して新たな数字タイルが出現する直前の盤面を *afterstate* と呼ぶ。またプレイヤーがゲームを開始する盤面を初期状態、ゲームが終了した盤面を終了状態と呼ぶ。図 2.4 に状態 s から *afterstate* s' を経由して、次の状態 s_{next} に遷移する例を示す。

プレイヤーの一般的な目標はゲームのタイトルが示す $2^{11} = 2048$ のタイルを完成させることだが、それ以降もゲームを続けることができる。

2.2 ゲームの進行上の性質

2048 はゲームの性質上、状態から *afterstate* への遷移において盤面上の数字タイルの合計値は不変である (図 2.1 を参照)。盤面上の数字タイルの合計値は *afterstate* から次の状態への遷移においてのみ変



図 2.1: 上下左右それぞれへのスライドの例

2	256	8	4
16	4	16	2
32	8	2	8
4	2	4	2

図 2.2: 終了状態の例

3	10	11	18
4	9	12	17
5	8	13	16
6	7	14	15

図 2.3: 理論的な最大到達盤面の例 (数字は指数部のみ)

化する．新しい数字タイルとして 2 か 4 のタイルが出現することで，数字タイルの合計値はその値の分だけ必ず増加する．すなわちプレイヤが 1 回行動するたびに，盤面上の数字タイルの合計値は 2 か 4 ずつ単調に増加する．

よって盤面上の数字タイルの合計値をゲームの進行度合いとして用いることができる．以降これを**時刻**

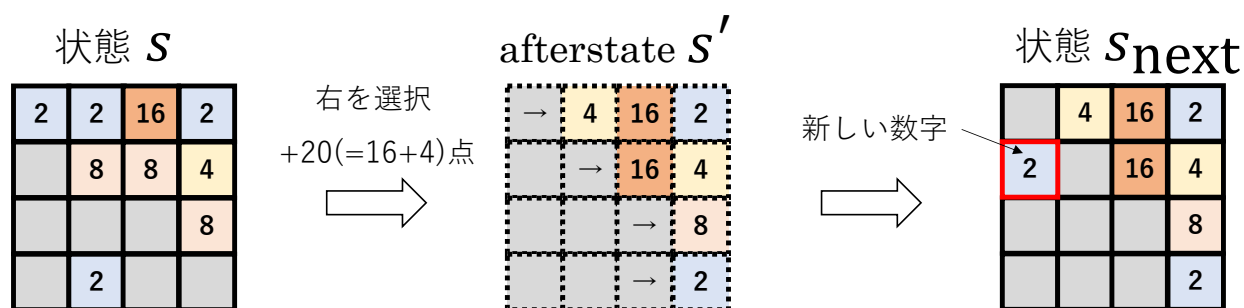


図 2.4: 状態遷移の例

と呼ぶ．例えば図 2.4 の遷移では時刻 $2 \times 4 + 4 + 8 \times 3 + 16 = 52$ の状態 s が時刻 $52 + 2 = 54$ の状態 s_{next} に遷移している．ゲームの時刻はプレイヤーが行動するたびに必ず増加するため，2048 はサイクルの出現しないゲームであることがわかる．

また 2048 は必ず終了するゲームである． 2^n のタイルを 1 つ完成させるには，盤面上に 2 つの 2^{n-1} のタイルを同時に存在させる必要がある．2 つの 2^{n-1} のタイルを同時に存在させるには，1 つの 2^{n-1} のタイルと 2 つの 2^{n-2} を同時に存在させる必要がある．新しい数字タイルとして 2 と 4 が出現することを踏まえると，帰納的に考えて 2^n のタイルを完成させるには最小でも $n - 2$ マスを必要とすることが分かる．よって 16 マスの 2048 では最高でも 2^{18} のタイルまでしか完成させられず，運や選択手に関わらずゲームを永久に続けることはできないことが分かる．

第 3 章

2048 と強化学習

これまでに 2048 を対象とした強化学習の研究は数多くなされてきた。本章では強化学習の概要、および 2048 に対する強化学習の先行研究について記述する。

3.1 強化学習の概要

まず本節では 2048 との関係を踏まえつつ、一般的な強化学習の概要について記述する。なお本節の内容は全体に文献 [2] を参照して書かれた。

3.1.1 マルコフ決定過程

強化学習は与えられた環境において試行錯誤することを通して、目標を達成するための戦略や意思決定を学習するための手法である。学習や意思決定を行う主体はエージェントと呼ばれる。エージェントは離散タイムステップに従って行動を選択し続け、環境とやり取りを行う。

このような問題設定はマルコフ決定過程 (MDP) というモデルによって定式化されている。MDP は以下の 4 つの要素で構成される。

- 状態集合 \mathcal{S}
- 行動集合 \mathcal{A}
- 状態遷移関数 $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$
- 報酬関数 $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$

エージェントはステップ t で状態 $S_t \in \mathcal{S}$ から行動 $A_t \in \mathcal{A}$ を選択する。そして確率 $p(S_{t+1}|S_t, A_t)$ で次の状態 S_{t+1} に遷移し、 $R_{t+1} = r(S_t, A_t, S_{t+1})$ を即時報酬として獲得する。状態遷移関数と報酬関数は環境のダイナミクスと呼ばれることがある。図 3.1 に MDP の概念図を示す。

状態集合と行動集合が有限である MDP を有限 MDP と呼ぶ。2048 は有限 MDP にそのまま当てはま

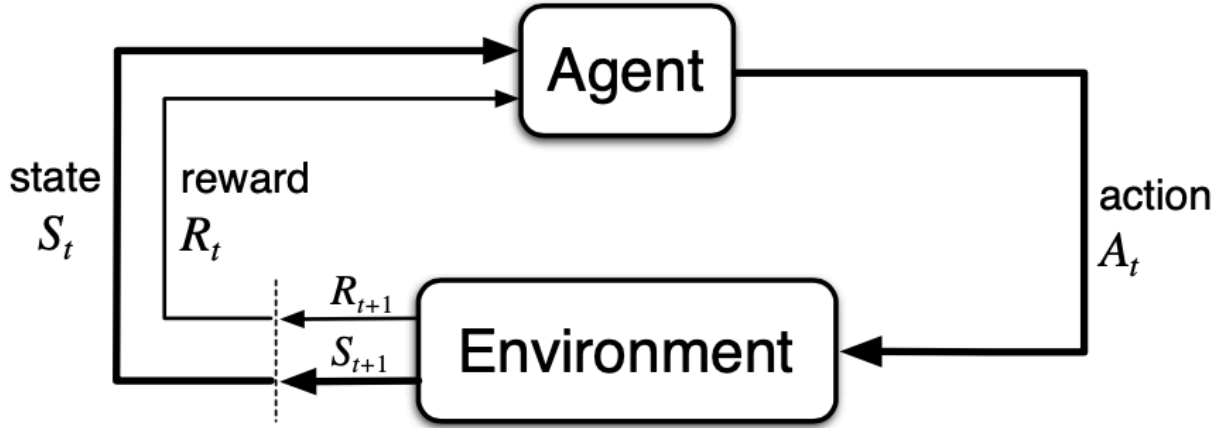


図 3.1: MDP の模式図 (文献 [2] より引用)

るゲームである。行動集合 A はプレイが選ぶ上下左右に対応し、報酬はプレイが獲得する得点に直接対応する。

一般に強化学習で扱う問題には、エージェントと環境のやり取りが終わる終了状態が存在する episodic task と終了状態が存在しない continuing task が存在する。episodic task ではエージェントと環境のやり取りを初期状態から終了状態までのエピソードと呼ばれる単位で分割することができる。2.2 節で説明したように 2048 は必ず終了するゲームであるため、以降 episodic task での定義を確認する。

3.1.2 方策と価値関数

エージェントがある状態において行動を決定する際の戦略、すなわち確率分布 $\pi : S \times A \rightarrow [0, 1]$ を方策と呼ぶ。状態価値関数 $v_\pi(s)$ は状態 s から方策 π に従って行動を選択し続けた場合の累積報酬和の期待値であり、次のように定義される。

$$v_\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\sum_{k=0}^T \gamma^k R_{t+k+1} | S_t = s \right] \quad (3.1)$$

同様に状態 s から行動 a を選択し、その後方策 π に従って行動を選択し続けた場合の累積報酬和の期待値である行動価値関数 $q_\pi(s, a)$ の定義は以下のようになる。

$$q_\pi(s, a) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\sum_{k=0}^T \gamma^k R_{t+k+1} | S_t = s, A_t = a \right] \quad (3.2)$$

強化学習の目標は多くの報酬を獲得できるような良い方策を見つけることである。価値関数の定義より 2 つの方策 π と π' があるとすると、すべての状態 $s \in S$ について $v_\pi(s) \geq v_{\pi'}(s)$ が成り立つならば π は π' と等価か π' よりも良い方策だと言える。ここで他のすべての方策と比べて等価であるか、それよ

りも良い方策が少なくとも 1 つ存在する．これは最適方策 π_* と呼ばれる方策である． π_* に従うときの状態価値関数は最適状態価値関数と呼ばれ， v_* で表される．同様に π_* に従うときの行動価値関数は最適行動価値関数と呼ばれ， q_* で表される．それぞれの具体的な定義を式 3.3, 3.4 に示す．

$$v_*(s) = \max_{\pi} v_{\pi}(s) \quad \text{for all } s \in S \quad (3.3)$$

$$q_*(s, a) = \max_{\pi} q_{\pi}(s, a) \quad \text{for all } s \in S \text{ and } a \in A(s) \quad (3.4)$$

このとき $v_*(s) = \max_{a \in A(s)} q_*(s, a)$ であるから，以下の式が導かれる (図 3.2 を参照) ．

$$v_*(s) = \max_{a \in A(s)} q_*(s, a) \quad (3.5)$$

$$= \max_{a \in A(s)} \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \quad (3.6)$$

$$q_*(s, a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \quad (3.7)$$

$$= \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') | S_t = s, A_t = a] \quad (3.8)$$

式 3.6, 3.8 はベルマン最適方程式と呼ばれる．

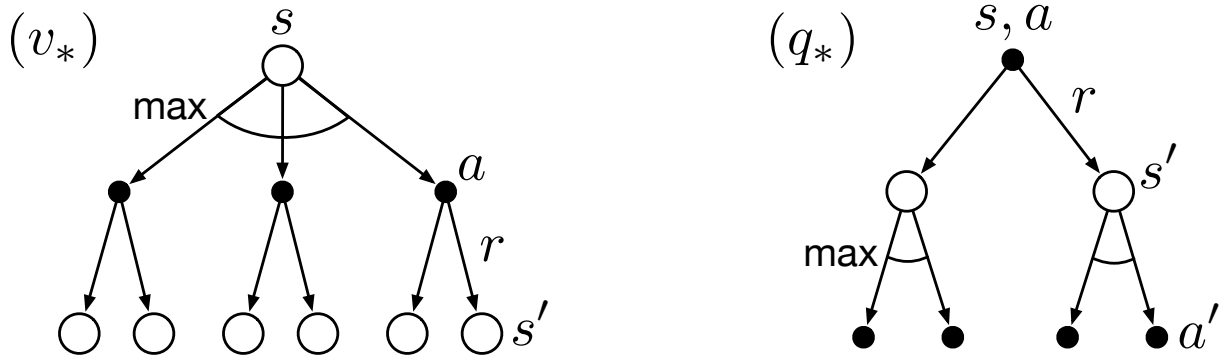


図 3.2: 最適価値関数のバックアップ図 (文献 [2] より引用)

3.1.3 価値ベースな手法

状態 s における最適方策に従った具体的な行動は $\arg \max \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$ と，1 ステップ先の状態を探索してそれらの最適状態価値関数を参照することで計算できる．最適行動価値関数が参照できれば，1 ステップ先の状態を探索する必要すらなく，単に $\arg \max q_*(s, a)$ を選択すればよい．最適方策を得るために価値関数を学習する手法は価値ベースな手法と呼ばれる．

Q 学習は最も基本的な価値ベースな手法の 1 つで，最適行動価値関数 q_* の推定値 Q を得られた経験から学習する．状態 S_t から行動 A_t を選択し，報酬 R_{t+1} を獲得して次の状態 S_{t+1} に遷移したとする．

このとき Q 学習は以下の更新式 3.9 に従って $Q(S_t, A_t)$ を更新する.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \quad (3.9)$$

式 3.9 は現在の推定値 $Q(S_t, A_t)$ を, ベルマン最適方程式 3.8 の右辺の推定値である $R_{t+1} + \gamma \max_a Q(S_{t+1}, a)$ に近づけていると解釈できる. 任意の $(s, a) \in \mathcal{S} \times \mathcal{A}$ について $Q(s, a)$ が更新され続けるという条件の下で, Q 学習は収束が保証されている.

状態集合や行動集合が大きい場合や有限でない場合には, テーブル形式で Q 値を保持することができないため何らかの関数近似を行う必要がある. 近年では深層学習の研究の発展により, 関数近似の方法としてニューラルネットワークが用いられることが多い. 価値関数や方策をニューラルネットワークで近似する手法は, まとめて深層強化学習と呼ばれる. Deep Q Network (DQN) [3] は深層強化学習の先駆けとして有名な手法である. 名前が示す通り DQN は Q 値をニューラルネットワークで近似し手法である. DQN の発表以降, 様々な工夫が提案され続けている. 詳細は文献 [4] の 4 節などを参照されたい.

3.1.4 方策ベースな手法

3.2 AlphaZero

Silver らが提案した AlphaZero [5] は, 二人ゼロ和完全確定情報ゲームを対象とした有力な深層強化学習手法である. 囲碁, 将棋, チェスにおいて当時の有力なプログラムを上回る強さを示した. AlphaZero は盤面の特徴量を入力として方策と価値を出力するニューラルネットワークを, 自己対戦を通して得たデータから学習する. 自己対戦において, AlphaZero はモンテカルロ木探索 (MCTS) というアルゴリズムを使用して指し手を選択する.

3.2.1 モンテカルロ木探索

ここでは AlphaZero が使用したモンテカルロ木探索 (MCTS) について説明する. MCTS は探索で見つけた各状態 (囲碁や将棋では盤面) に対応するノードから成る探索木を構築する. それぞれの状態と行動の対 (s, a) について, $N(s, a), W(s, a), Q(s, a), P(s, a)$ という 4 つの統計量を管理する.

- $N(s, a) \cdots$ 探索中に s から a を選択した回数
- $W(s, a) \cdots$ 最適行動価値 $q_*(s, a)$ の推定値の累計
- $Q(s, a) \cdots$ 最適行動価値 $q_*(s, a)$ の推定値の平均 ($W(s, a)/N(s, a)$)
- $P(s, a) \cdots$ ニューラルネットワークが評価する s から a を選択する確率

探索木は最初, 現在の盤面に対応するノードである根ノードのみから成る. MCTS は選択, 展開, 評価, 逆伝播の 4 つのステップを繰り返すことで, 良い手を選択するための探索手法である. この 4 つの

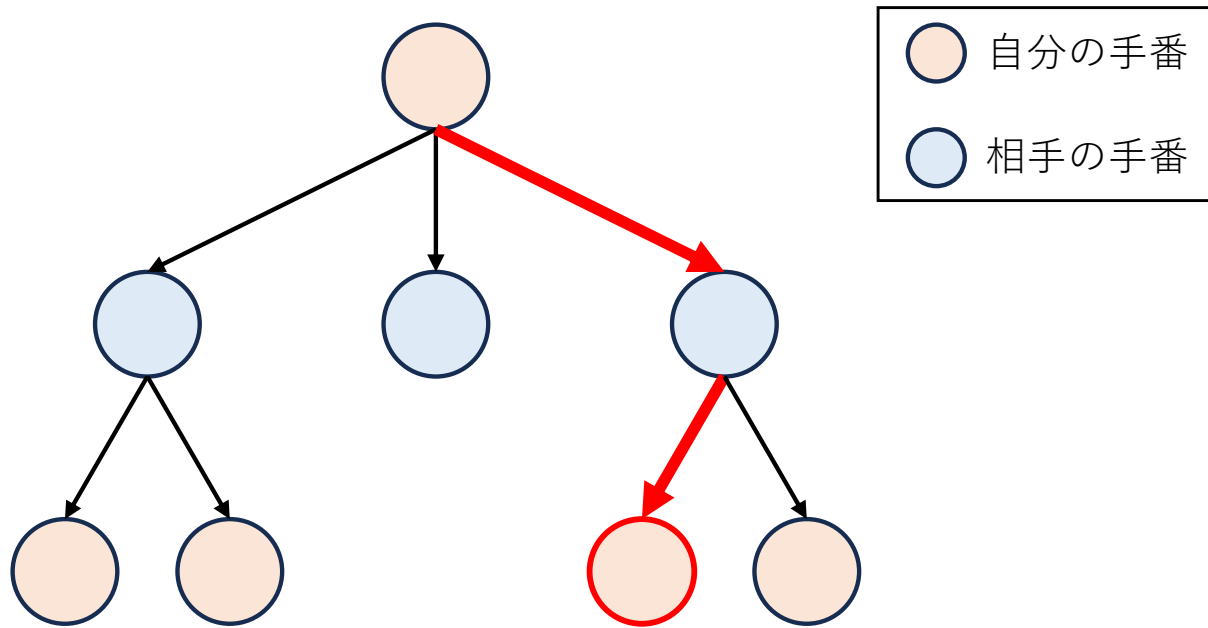


図 3.3: 選択

ステップはまとめてシミュレーションと呼ばれる。シミュレーションを繰り返すことで探索木は深くなり、良い手を選択することができる。以下ではそれぞれのステップについて説明する。

選択

現在のゲーム木の根ノードから有望な子ノードを選択し、葉ノードに至るまで辿り続ける。ここで有望なノードとは暫定の評価の高さと

展開

評価

逆伝播

3.2.2 ニューラルネットワークの学習

MCTS の探索木における各ノードは MCTS は選択、展開、評価、逆伝播という 4 つのステップを繰り返すことで良い手を選ぶための手法である。

3.3 2048 への強化学習の応用

3.1 節では強化学習一般の概要について説明した。これを踏まえて本節では、2048 を対象とした強化学習研究の先行研究について概説する。

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \quad (3.10)$$

3.3.1 TD afterstate 学習

Szuber ら [6] は TD afterstate 学習と呼ばれる価値ベースの学習手法を提案した。TD afterstate 学習では afterstate の価値 N-tuple ネットワークは 2048 の盤面特徴量の提案した TD afterstate 学習は現在でも主流の手法の 1 つである。

3.3.2 Afterstate PPO

3.3.3 Stochastic MuZero

第 4 章

2048 と完全解析

3 章で述べた強化学習は環境 (ゲーム) と何度もやり取りすることで、最適な方策を学習するための手法である。一方で小さなゲームであれば、力ずくの計算によってゲームを完全に解くこともできる。本章では 2048 を解析的なアプローチによって解くことについて考察する。

4.1 2048 の完全解析とは

2048 は 1 人用のゲームであるため、勝敗のようなプレイヤーの明確な目標は存在しない。そのためプレイヤーが何を目標とするかによって、プレイヤーの最善手の定義は変化する。また 2.1 節で述べたようにゲームはランダム性を伴うため、同じ状態から毎回同じ手を選んでも結果は確率的に変動する。

そこで本稿ではある状態 s における最善手を「 s から獲得できる得点の合計の期待値が最も高くなるような手」と定義する。これは 3 節で述べた強化学習の最適状態価値と同じものである。本節では状態 s から最善手を選び続けて獲得できる得点の合計の期待値を状態 s の最適価値と呼び、 $v(s)$ で表すことにする。このとき $v(s)$ は式 4.1 のように再帰的な形式で書くことができる。ただし $r(s, a)$ は状態 s から行動 a をとって獲得する得点、 $s_{\text{next}} \in \mathcal{T}(s, a)$ は状態 s から行動 a をとって遷移しうる次の状態の集合を表す (図 4.1 を参照)。

$$v(s) = \begin{cases} 0 & (s \text{ が終了状態}) \\ \max_a (r(s, a) + \mathbb{E}_{s_{\text{next}} \in \mathcal{T}(s, a)} v(s_{\text{next}})) & (\text{otherwise}) \end{cases} \quad (4.1)$$

ゲームに現れうるすべての状態の価値を計算すれば、任意の状態において最善手を選ぶことができる。本稿ではこれを 2048 の完全解析ということにする。

完全解析をすることで、最善手を選び続けるプレイヤーの戦略やゲームの初期状態の価値を知りたい人は少なくないだろう。また 2048 を対象とした強化学習手法の良し悪しを定量的に評価することができる。一方で、2048 を完全解析することはそのゲーム木の大きさから現状難しいと考えられる。そこで以降では本来 4×4 盤面上で行われる 2048 のミニゲームとして、盤面サイズを縮小した 2048 を完全解析

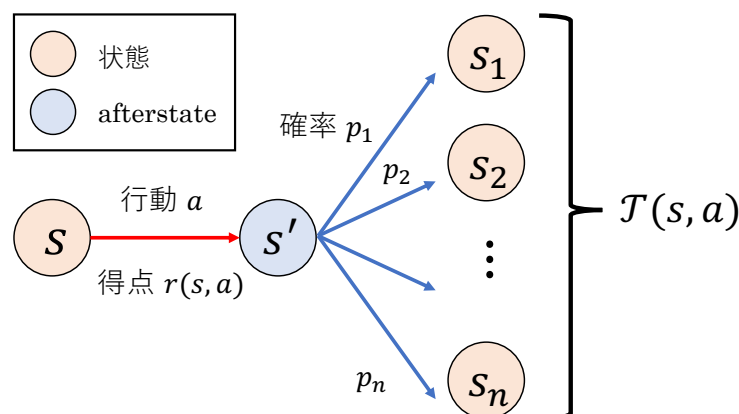


図 4.1: 式 4.1 の補足図

することを考える。

4.2 盤面サイズが小さな 2048 の完全解析

基本的なルールは 2048 と同じで盤面サイズを 4×4 から縮小したゲームを完全解析することを考える。盤面サイズに関わらず、以下の 2 つのステップを順番に行うことで 2048 を完全解析することができる。

1. ゲームに現れうるすべての状態の列挙
2. 列挙した状態の価値の計算

4.2.1 節と 4.2.2 節で具体的な方法について述べる。なお本節の内容は文献 [7] および文献 [8] を元に執筆された。

4.2.1 幅優先探索によるすべての状態の列挙

完全解析の第 1 ステップとして幅優先探索によってゲームに現れうるすべての状態を列挙する。まず初期状態をキューに詰めて探索を開始する。キューの先頭の状態 s を取り出し、 s から遷移可能な次の状態 $s_{\text{next}} \in \mathcal{T}(s)$ をキューに追加する。これをキューが空になるまで繰り返すことで、すべての状態を列挙することができる。

ここで $s_{\text{next}} \in \mathcal{T}(s)$ がすでに発見済みであるか確認するために、これまでに発見した状態を管理する集合が必要である。素朴な方法ではメモリでこれまでに発見した全状態を管理することで行えるが、状態数が非常に大きな場合にはメモリの容量を超えてしまう。

そこで 2.2 節で説明した時刻によってゲーム木を整理する。時刻 t の状態は時刻 $t+2$ か $t+4$ の状態にしか遷移しないため、時刻 $t+2$ と $t+4$ の発見した状態をメモリで管理すれば十分である。よって時

Algorithm 1 幅優先探索によるすべての状態の列挙

```

1: function ENUMERATION( $t$ )
2:   for all  $s_t \in \text{queue}_t$  do
3:     for all  $s_{t+2} \leftarrow s_t$  do
4:        $\text{queue}_{t+2}.\text{push}(s_{t+2})$ 
5:     end for
6:     if  $\text{element} > \text{max}$  then
7:        $\text{max} \leftarrow \text{element}$ 
8:     end if
9:   end for
10:  return  $\text{max}$ 
11: end function

```

Algorithm 2 後退解析による価値計算

```

1: function CALCULATION( $t$ )
2:   for all  $s_t \in \text{queue}_t$  do
3:     for all  $s_{t+2} \leftarrow s_t$  do
4:        $\text{queue}_{t+2}.\text{push}(s_{t+2})$ 
5:     end for
6:     if  $\text{element} > \text{max}$  then
7:        $\text{max} \leftarrow \text{element}$ 
8:     end if
9:   end for
10:  return  $\text{max}$ 
11: end function

```

刻が最小の 4 の状態から時刻 2 刻みで順番に列挙を行うことで、ディスクを効率的に活用することができる。以上を踏まえた疑似コードを Algorithm 1 に示す。

4.2.2 後退解析による状態の価値の計算

4.2.1 節で列挙した状態の価値を式 4.1 に従って計算する。時刻 t の状態の価値は、時刻 $t+2$ と $t+4$ の状態の価値が計算済みであれば計算できる。よって時刻が最大の状態から順番に走査することで、効率的にすべての状態の価値を計算できる。

4.2.3 実験結果

第 5 章

提案手法

これまでに 2048 を対象とした強化学習の研究は数多くなされてきた．ゲームを完全解析

5.1 3×3 盤面の 2048 の完全解析と強化学習

付録 A

実装の詳細

A.1 ゲーム環境の実装

2048 は状態から afterstate への遷移において、各行 (列) の変化は独立に考えることができる。また回転と反転を考慮することで上下左右は等価な盤面変化を起こす。よって 1 行の全パターンについて、ある一方向を選択したときの遷移先を前もって計算することで、全方向に対する盤面全体の遷移を高速に行える。

A.2 完全解析の実装

A.3 強化学習の実装

参考文献

- [1] G Cirulli. 2048, available from <http://gabrielecirulli.github.io/2048/>, 2014.
- [2] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.
- [3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, Vol. 518, No. 7540, pp. 529–533, 2015.
- [4] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. *CoRR*, Vol. abs/1811.12560, , 2018.
- [5] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, Vol. 362, No. 6419, pp. 1140–1144, 2018.
- [6] Marcin Szubert and Wojciech Jaśkowski. Temporal difference learning of n-tuple networks for the game 2048. In *2014 IEEE CIG*, pp. 1–8, 2014.
- [7] 山下修平, 金子知適, 中屋敷太一. 3×3 盤面の 2048 の完全解析と強化学習の研究. ゲームプログラミングワークショップ 2022 論文集, 第 2022 巻, pp. 1–8, nov 2022.
- [8] 山下修平, 金子知適. 4×3 盤面の 2048 の完全解析. ゲームプログラミングワークショップ 2023 論文集, 第 2023 巻, pp. 1–5, nov 2023.