強化学習
Reinforcement Learning

津田塾大学　学芸学部　情報科学科
新田善久
nitta@tsuda.ac.jp

強化学習

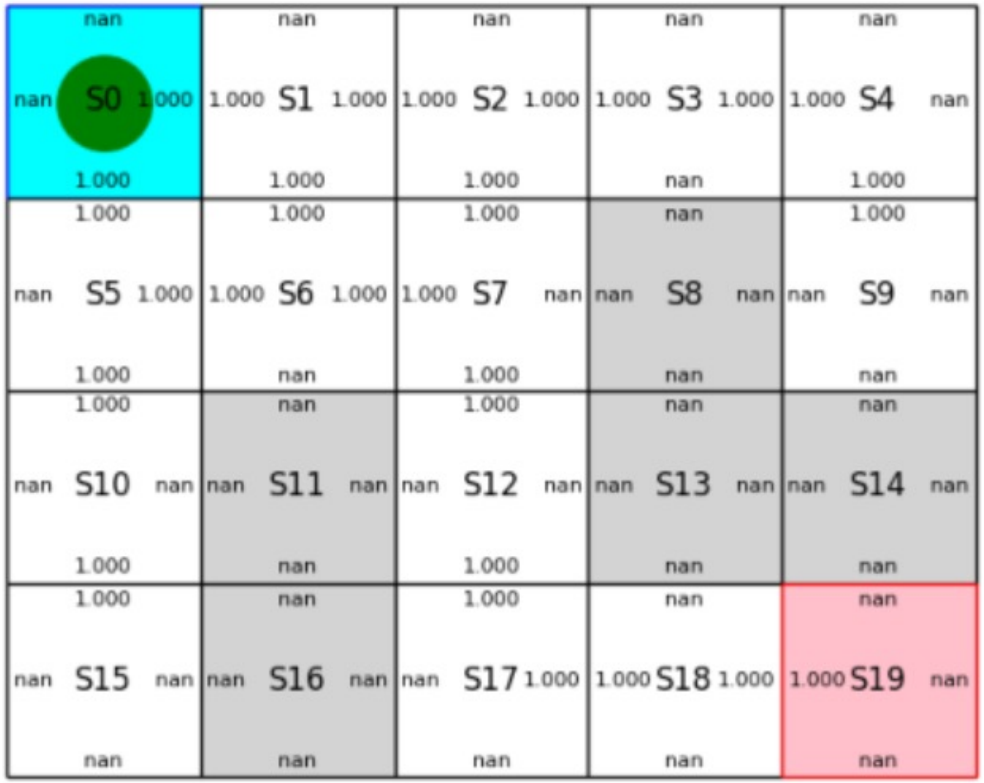
- 「対象を支配している法則を知らない状態で試行錯誤を繰り返しながら、望ましい制御方法を学習する」手法
- 何度も試行錯誤しながら、成功した場合は正の報酬を与え、失敗した場合は負の報酬を与えることで、システムは自身の報酬が最大化するように制御ルールを更新していく。
- その結果、各局面でシステムが望ましい行動を取るようになる。

深層強化学習

- 強化学習では、システムに「状態s(t)」を与えると、次に行うべき「行動a(t)」を出力する。
- ただし、すべての状態におけるすべての「行動」の組み合わせを列挙するには「状態のパターン数」×「行動の種類数」の大きさの表が必要になる。
- この表が大きくなり過ぎると、その問題を扱えなくなってしまう。
- 何らかの方法で「状態を圧縮して表現する」必要がある。
- そのために「深層ニューラルネットワーク」を利用するのが「深層強化学習」である。

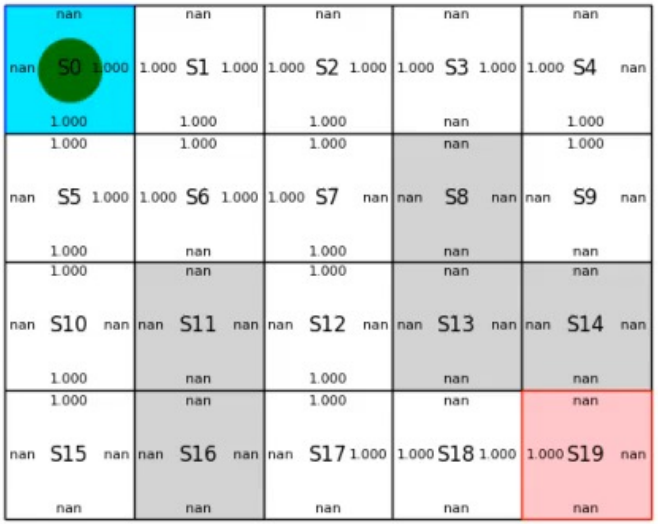
強化学習の例：迷路脱出

- 迷路 縦H×横Wのマス
- マスの意味
  - 水色：スタート地点
  - 赤色：ゴール
  - 白色：通過できる
  - 灰色：障害物
- 状態: どのマスにいるか
- 行動: 上下右左のいずれかに移動



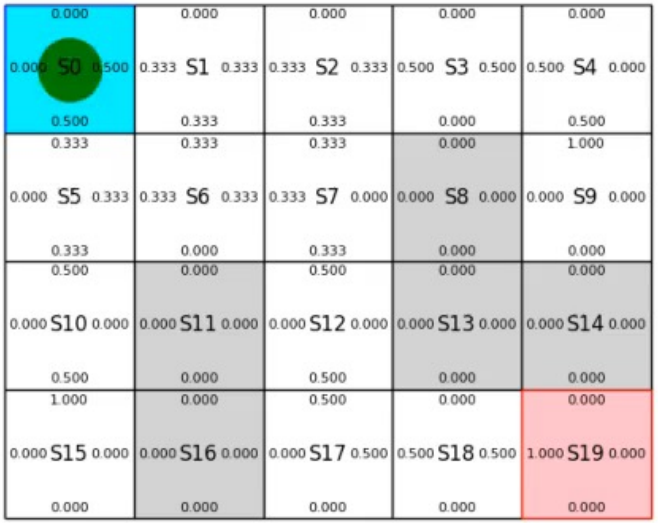
強化学習の例：迷路脱出

- ランダムに行動を選択する。



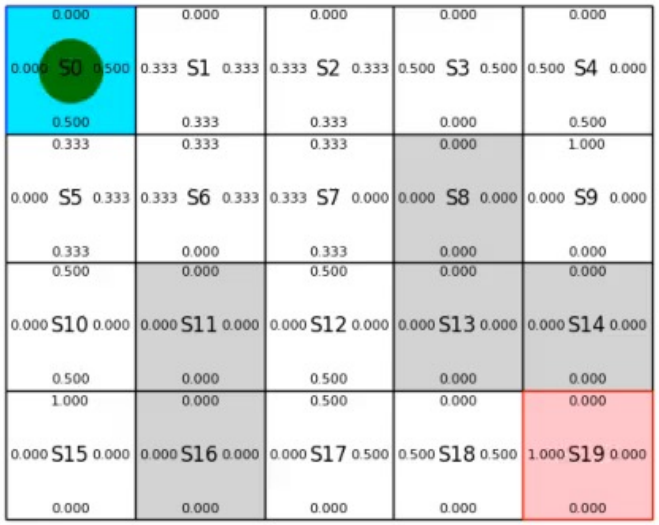
強化学習の例：迷路脱出

- 方策π_0の確率分布にしたがって、行動を選択する
- 各マスの数字は、その方向に移動する確率



強化学習の例：迷路脱出:方策勾配法

- 成功した一連の行動について、その中に現れた (状態, 行動) の重みを更新する。
- 重みにしたがった確率で行動を選択する。



強化学習の例：迷路脱出:価値反復法

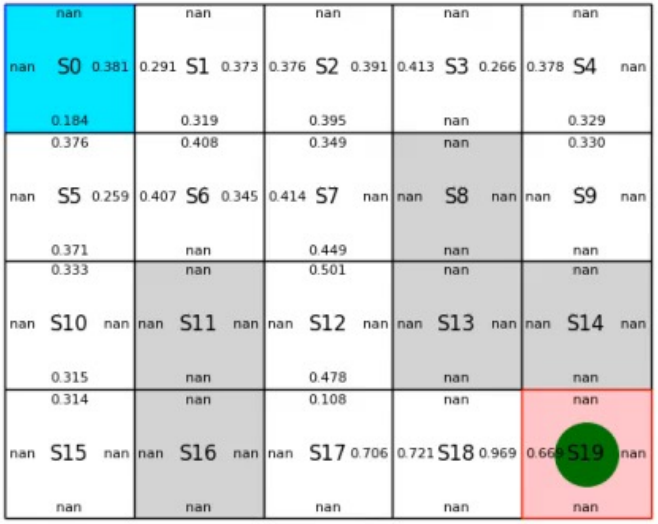
- 成功したときに、報酬が与えられる。
- 失敗したときは、ペナルティ（負の報酬）が与えられる。
- 時刻 t でもらえる報酬 R_t （即時報酬)
- 未来の報酬を γ ずつ割り引いて考えると
  割引報酬和

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

- 行動価値関数 Q(s,a) を表形式で実装する。
- 確率 ε でランダムに行動し、確率 1-ε で最大の Q となる行動を選択する (ε-greedy 法)

強化学習の例：迷路脱出:価値反復法: Sarsa アルゴリズム

- 行動価値関数 Q(s, a) を Sarsa アルゴリズムで更新する。



強化学習の例：迷路脱出:価値反復法:
Q 学習

- Q 学習は Sarsa とは行動価値関数 Q の更新式が異なる。
- Q 学習では、状態 st+1 における行動価値関数の最大値を使用して Q を更新する。

Sarsaの場合:

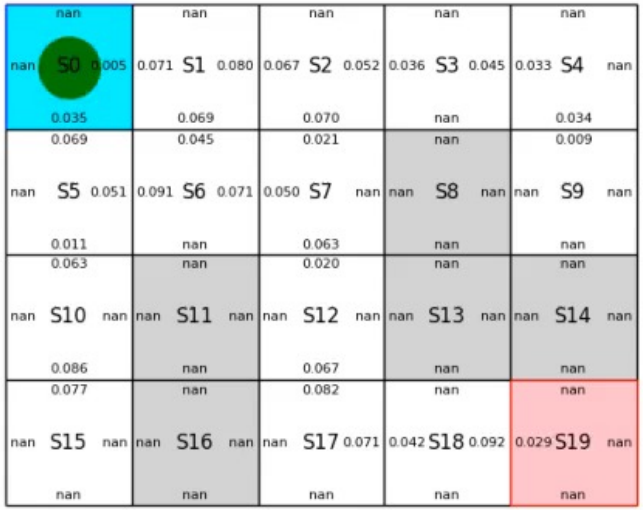
$$Q(S_t, a_t) = Q(S_t, a_t) + \tau * (R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

Q学習の場合:

$$Q(S_t, a_t) = Q(S_t, a_t) + \tau * (R_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t))$$

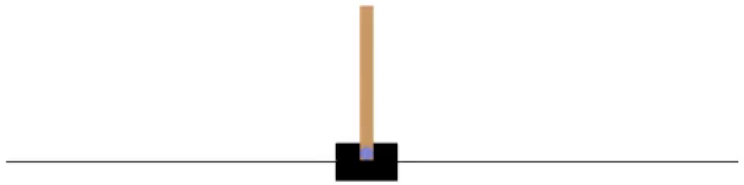
- Q 学習は乱数を使わないので行動価値関数の収束が早い。

強化学習の例：迷路脱出:価値反復法:
Q 学習



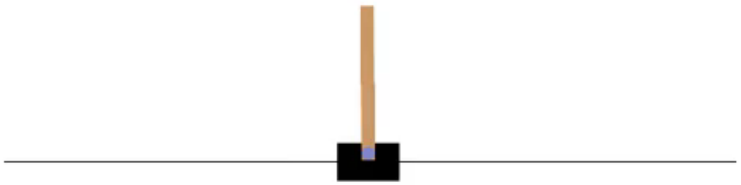
強化学習の例: CartPole(回転振子課題)

- 台車の上に回転軸を固定した棒を立てて、立てた棒が倒れないように台車を右・左に動かす制御問題。
- 小学生の頃、掃除の時間に、手のひらの上にほうきを立てて倒れないように遊んだ、あのゲーム。
- CartPoleの状態
  - カート位置 (-2.4, 2.4)
  - カート速度(-∞, +∞)
  - 棒の角度 (-41.8°, 41.8°)
  - 棒の角速度(-∞, +∞)
- 行動: 右 or 左 にカートを動かす



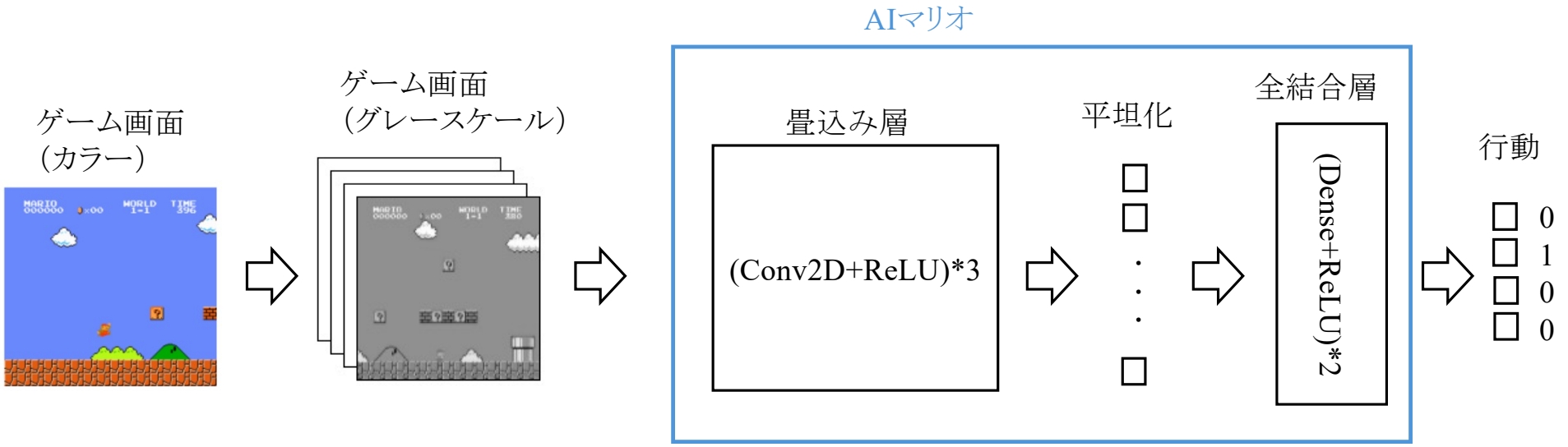
強化学習の例: CartPole: Double DQN

- 倒れないでいれば、報酬を与える
- Main と Target の2つの Q ネットワークを用意する。Target を利用して行動を決定し、学習結果はMain に反映させる。Target は時々 Main で上書きして更新する。(学習が安定する)



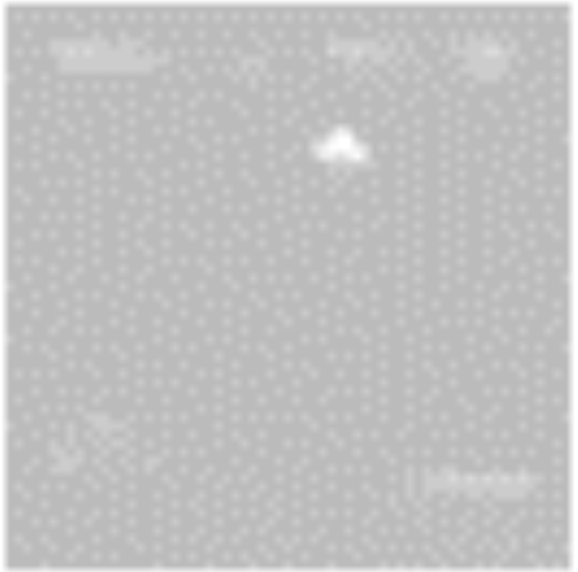
強化学習の例: スーパーマリオ

- ゲーム画面を84x84のグレースケール画像に変換して、4画面まとめてモデルに与え、行動(右,右A,右B,右AB)を出力する。



強化学習の例: スーパーマリオ

- 行動: 右/右ジャンプ/右高速/右高速ジャンプ
- 右に多く進むほど報酬を与える。
- 84x84のグレースケール画像に変換した画面を4画面ごとに与えて、行動を選択させる



強化学習の例: スーパーマリオ: 学習済み

- 40000 エピソードの学習を繰り返したところ、1-1面をクリアできるようになった。