計算知能
(COMPUTATIONAL INTELLIGENCE)

第１４回　強化学習

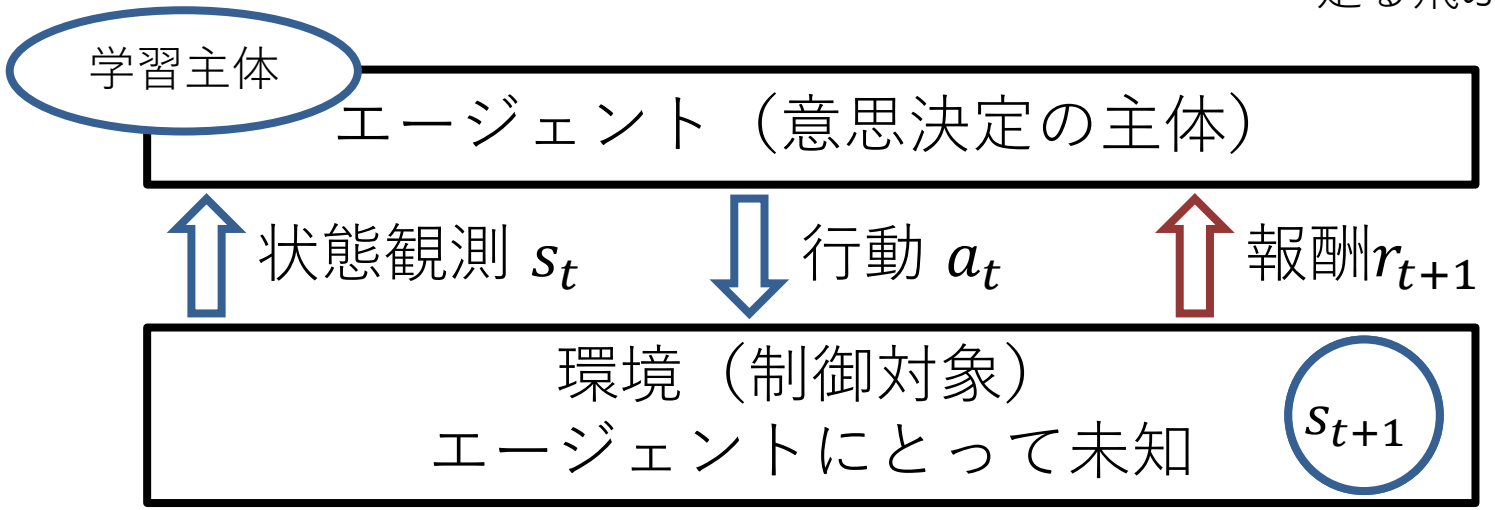
教員：　谷口彰

第14回 強化学習

- マルコフ決定過程
- Q-learning

強化学習とは(1/5)

- 試行錯誤を通じて環境に適応する学習制御の枠組み
- 生体の「脳」のシステムを模倣



- 教師あり学習とは異なり、状態入力に対する正しい行動出力を明示的に示す教師が存在しない

強化学習とは(2/5)

- 状態観測→行動選択→（状態遷移）→報酬　繰り返し
- 何回か状態遷移した後、ようやく最終的な報酬を得る
  → 多段決定過程（報酬に遅れ）
- 報酬合計が最大になる行動列を探索する

行動列：
走る→飛ぶ→手を放す

最初はstartから出発し、goalにたどり着いたときに報酬合計を獲得

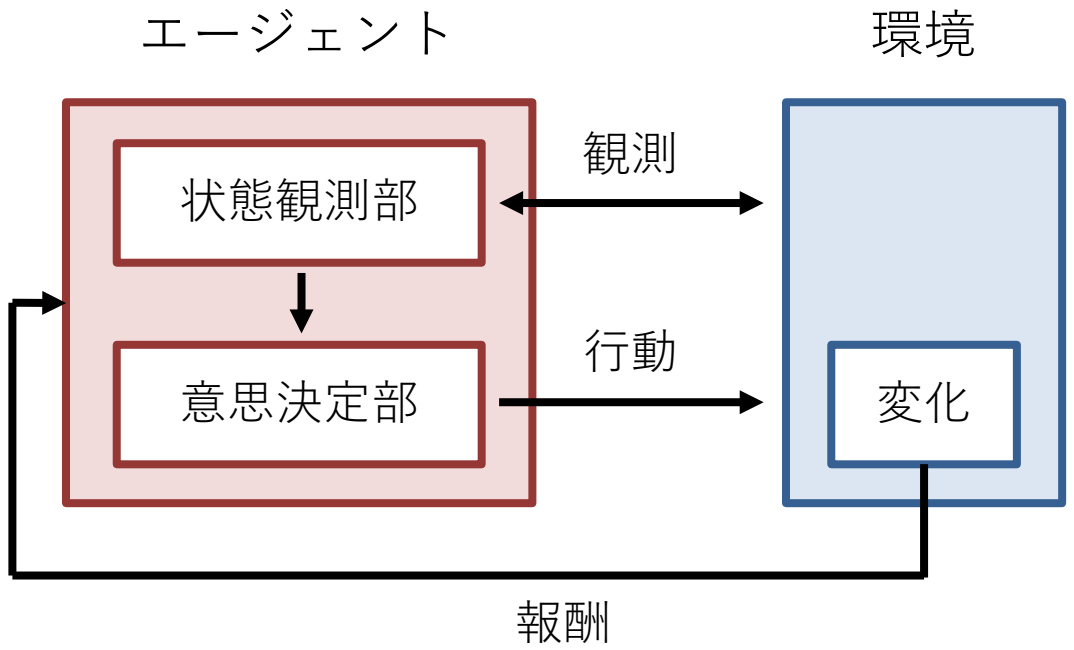
強化学習とは(3/5)

- 学習主体「エージェント」と制御対象「環境」は
  以下のやりとりを行う

1. エージェントは時刻tにおいて環境の状態観測s_tに応じて
   意志決定を行い、行動a_tを出力
2. エージェントの行動により、環境はs_{t+1}へ状態遷移し、
   その遷移に応じた報酬r_{t+1}をエージェントへ与える
3. 時刻tを$t+1$に進めてステップ1へ戻る

強化学習とは(4/5)

- エージェントは利得（return: 最も単純な場合、報酬の合計）の最大化を目的として、状態観測から行動出力へのマッピング（方策（policy） と呼ばれる）を獲得する



強化学習とは(5/5)

- 環境とエージェントの一般的な性質
  - エージェントは予め環境に関する知識を持たない
  - 環境の状態遷移は確率的
  - 報酬の与えられ方は確率的
  - 状態遷移を繰返した末に最終的な報酬を得るような、段階的な行動を必要とする環境（報酬の遅れ）

制御の視点から見た強化学習

- 強化学習が注目を集める理由：
  - 不確実性のある環境を扱っている点
  - 報酬に遅れが存在し、離散的な状態遷移も含んだ段階的な制御規則の獲得を行う点

強化学習では、設計者がゴール状態で報酬を与えるという形で、させたいタスクをエージェントに指示しておけば、ゴールへの到達方法はエージェントの試行錯誤学習によって自動的に獲得される

制御プログラミングの自動化・省力化

- 環境に不確実性や計測不能な未知のパラメータが存在すると、タスクの達成方法やゴールへの到達方法は設計者にとって自明ではない
- 制御規則をプログラムすることは設計者にとって重労働

- 達成すべき目標を報酬によって指示することは上記に比べ遥かに簡単

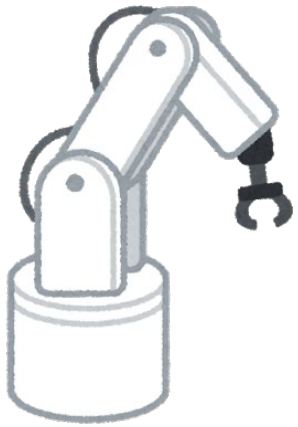
マルコフ決定過程(1/4)

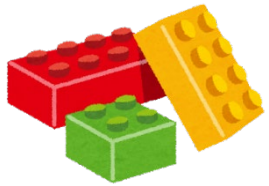
- 環境のとりうる状態の集合を

$$\mathbf{s} = \{s_1, s_2, \cdots, s_n\}$$

エージェントがとりうる行動の集合を

$$\mathbf{a} = \{a_1, a_2, \cdots, a_m\}$$





マルコフ決定過程(2/4)

ある状態 $s \in \mathbf{s}$において、エージェントがある行動 a を
実行すると、環境は確率的に状態 $s' \in \mathbf{s}$へ遷移する

遷移確率： $P\{s_{t+1} = s' | s_t = s, a_t = a\} = P^a(s, s')$

環境からエージェントへ報酬rが与えられる期待値：

$$E\{r_t | s_t = s, a_t = a, s_{t+1} = s'\} = R^a(s, s')$$

マルコフ決定過程(3/4)

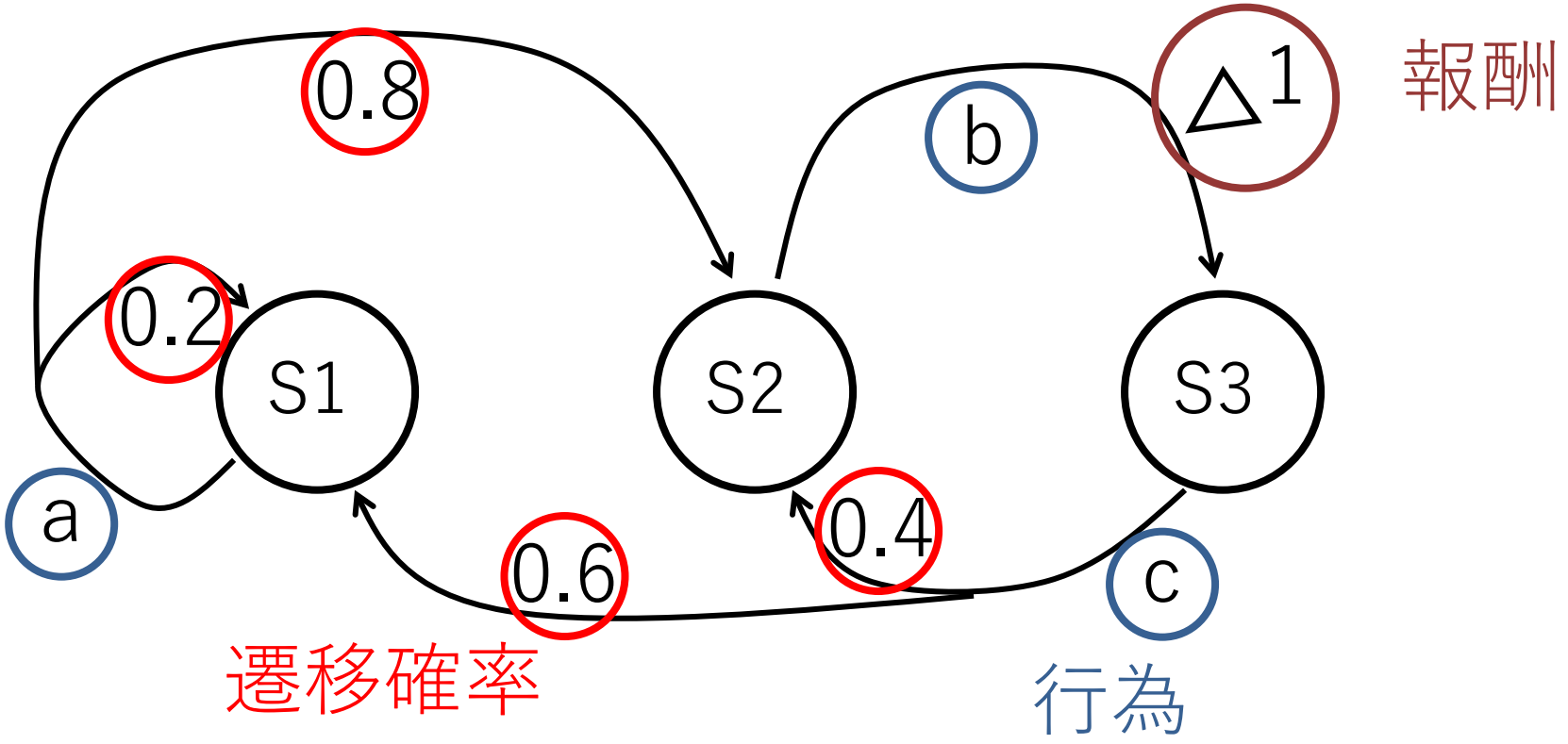
- エージェントの各時刻における意志決定は、
  方策$\pi(s,a) = P\{a_t = a | s_t = s\}$（ただし全状態$s$、全行動$a$において定義される）によって表される

2つの性質

- マルコフ性： 状態s'への遷移が、そのときの状態sと行動aにのみ依存し、それ以前の状態や行動には関連しない
- エルゴート性： 任意の状態sからスタートし、無限時間経過した後の状態分布確率は 最初の状態とは無関係

マルコフ決定過程(4/4)

- 状態遷移確率は現在の状態のみに依存する
- 状態遷移確率は時刻によって変化しない



MDPの最適性：割引報酬による評価(1/2)

- ある時刻tで実行した行動が、その後の報酬獲得にどの程度貢献したのかを評価するには？

  → その後得られる報酬の時系列を考えれば良い

利得 $V_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$

r_t：時刻tにおける報酬

- 時間経過とともに報酬を割引率 $\gamma (0 \leq \gamma < 1)$ で割引いて合計

MDPの最適性：割引報酬による評価(2/2)

- 実環境では時間の経過とともに環境が変化し、エージェントが故障等で停止する可能性がある

⇩

時系列上の報酬の重みを変化

- マルコフ決定過程においてエージェントがある方策πをとるとき、利得の期待値は状態sに依存して決まる。

  状態価値関数と呼び、$V^{\pi}(s)$と表す

$$\text{状態価値関数} v(s) := E[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) | s_0 = s]$$

最適な状態価値関数

- 全ての状態sにおいて$V^{\pi}(s) \geq V^{\pi'}(s)$となるとき、
  方策πはπ'より優れているという

- マルコフ決定過程では、他のどんな方策よりも優れた
  （あるいは同等な）方策が少なくとも1つ存在（最適方策 π^*）

$$V^*(s) = \max_{\pi} V^{\pi}(s) \quad \text{for all } s \in \mathbf{s}$$

最適な行動価値関数

- 状態価値関数は、行動価値関数と方策で表現できる

$$v(s) = \sum_a \pi(a|s) Q(s,a)$$

- 最適な方策π^*は以下の唯一の行動価値関数を共有する

$$Q^*(s,a) = \max_\pi Q^\pi(s,a) \quad \text{for all } s \in \mathbf{s} \text{ and } a \in \boldsymbol{a}$$

- $Q^*(s,a)$はQ値と呼ばれ、状態sで行動aを選択後、常に最適方策とりつづけた際の利得の期待値
- $Q^*(s,a)$が与えられた場合、状態sにおいて最大のQ値を持つ行動aが最適な行動

マルコフ決定過程における強化学習

- マルコフ決定過程での強化学習問題は以下の様に定式化される
  - エージェントは環境の状態遷移確率$P^a(s, s')$や
    報酬の与えられ方$R^a(s, s')$についての知識を予め持たない
  - エージェントは環境との試行錯誤的な相互作用を
    繰り返して最適な方策を学習する

- $Q^*(s, a)$が得られれば最適な方策は簡単に得られる

Q-learningの処理手順(1/2)

状態と行動の二次元配列の値を以下のように環境とのインタラクションに応じて変数を修正

1. 状態と行動の2次元配列$Q(s, a)$を用意して0に初期化する
2. 環境の状態s_tを観測する
3. 行動選択方法（探査戦略）に従って行動aを実行する
4. 遷移後の状態s_{t+1}を観測する

Q-learningの処理手順(2/2)

4. 状態s_{t+1}において報酬r_{t+1}を獲得する
5. 更新式によりQ値を更新する（α：学習率，γ：割引率）

$$Q(s_t, a) \leftarrow (1-\alpha)Q(s_t, a) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right]$$

6. 時刻 t を $t+1$ へ進めて手順２に戻る

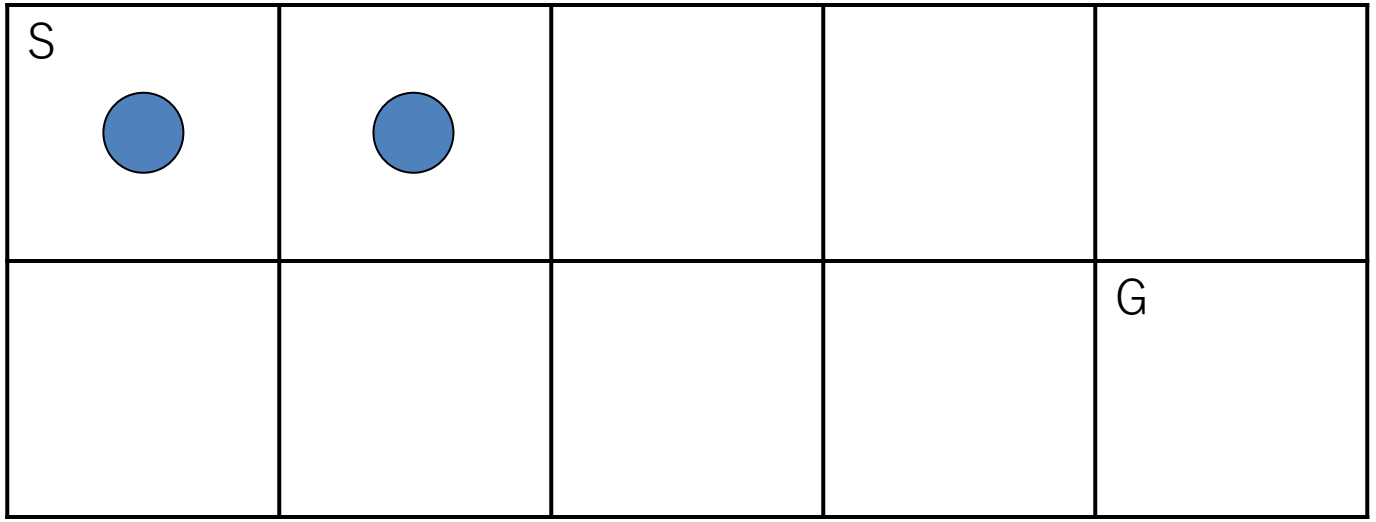
行動a：(up, down, left, right), 報酬r：(G：1, 他：0)

S 1	2	3	4	5
6	7	8	9	G 1 0

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1\text{か}0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

行動a：(up, down, left, right), 報酬r：(G：1, 他：0)

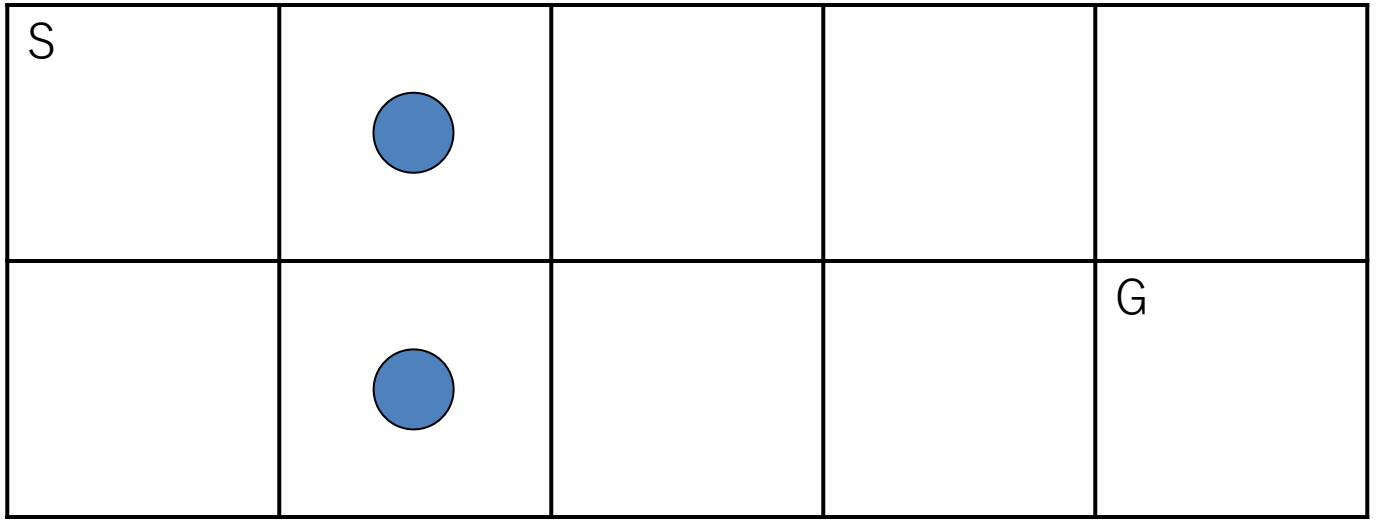


$$Q(1, right) \leftarrow 0 + 0.5[0 + 0.1 \times 0]$$

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1\text{か}0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

行動a：(up, down, left, right), 報酬r：(G：1, 他：0)

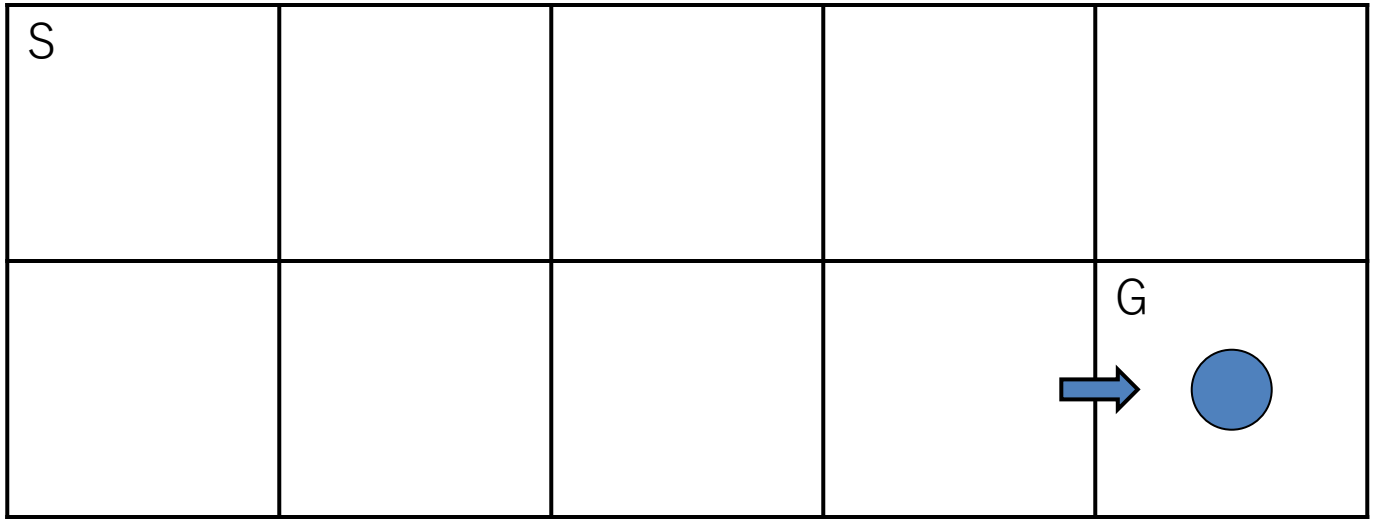


$$Q(2, down) \leftarrow 0 + 0.5[0 + 0.1 \times 0]$$

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1\text{か}0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

行動a：(up, down, left, right), 報酬r：(G：1, 他：0)

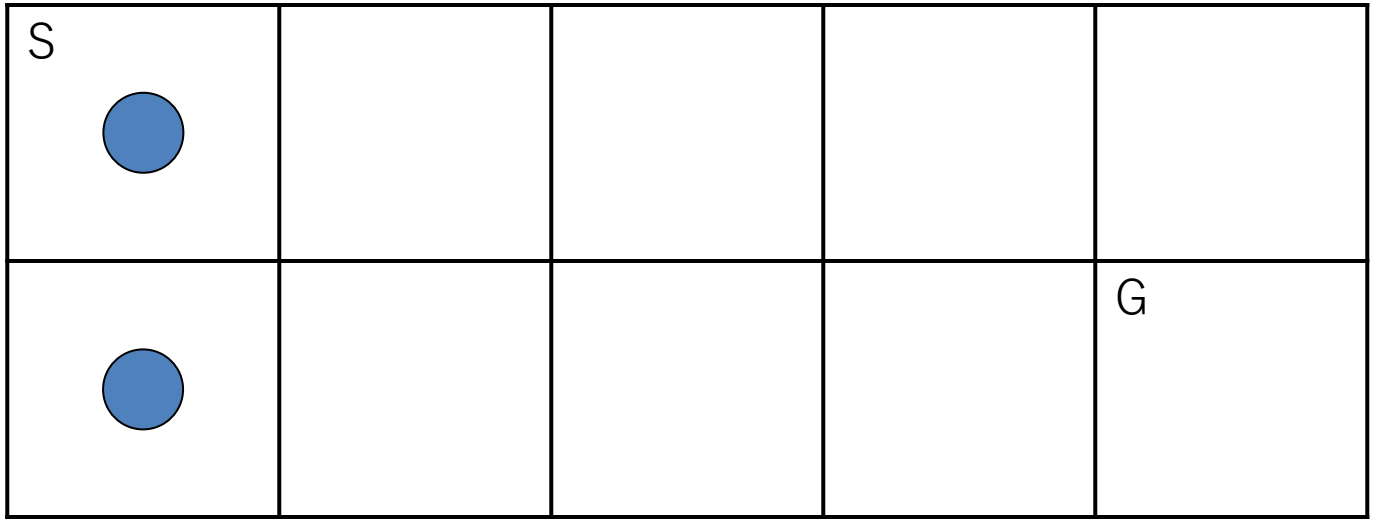


$$Q(9, right) \leftarrow 0 + 0.5[1 + 0.1 \times 0]$$

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1か0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

行動a：(up, down, left, right), 報酬r：(G：1, 他：0)

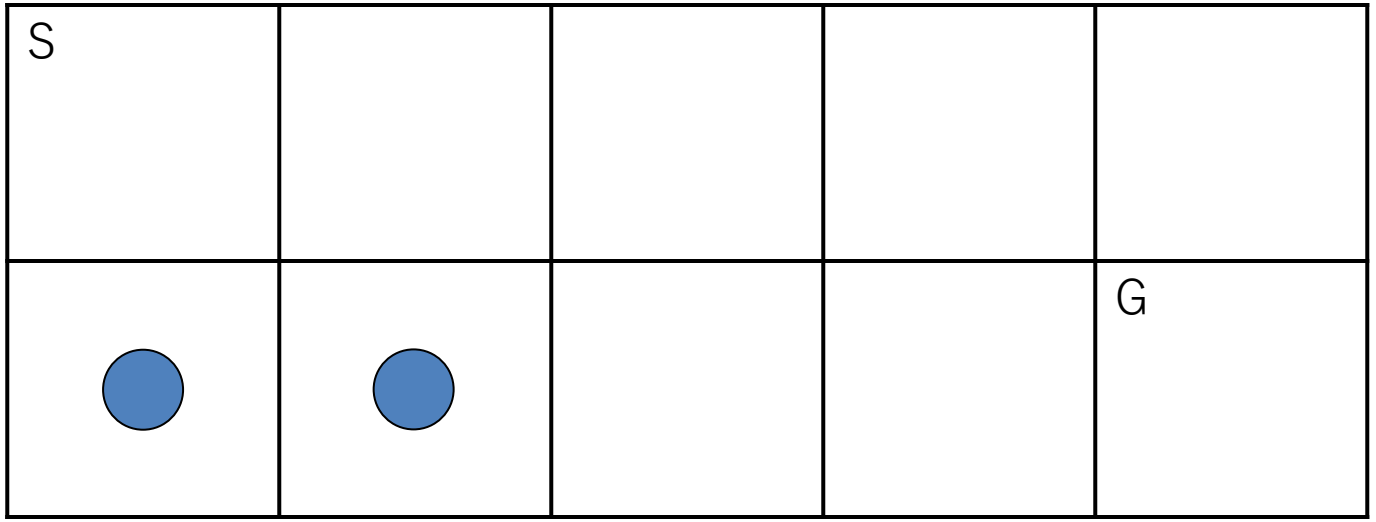


$$Q(1, down) \leftarrow 0 + 0.5[0 + 0.1 \times 0]$$

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1か0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

行動a：(up, down, left, right), 報酬r：(G：1, 他：0)

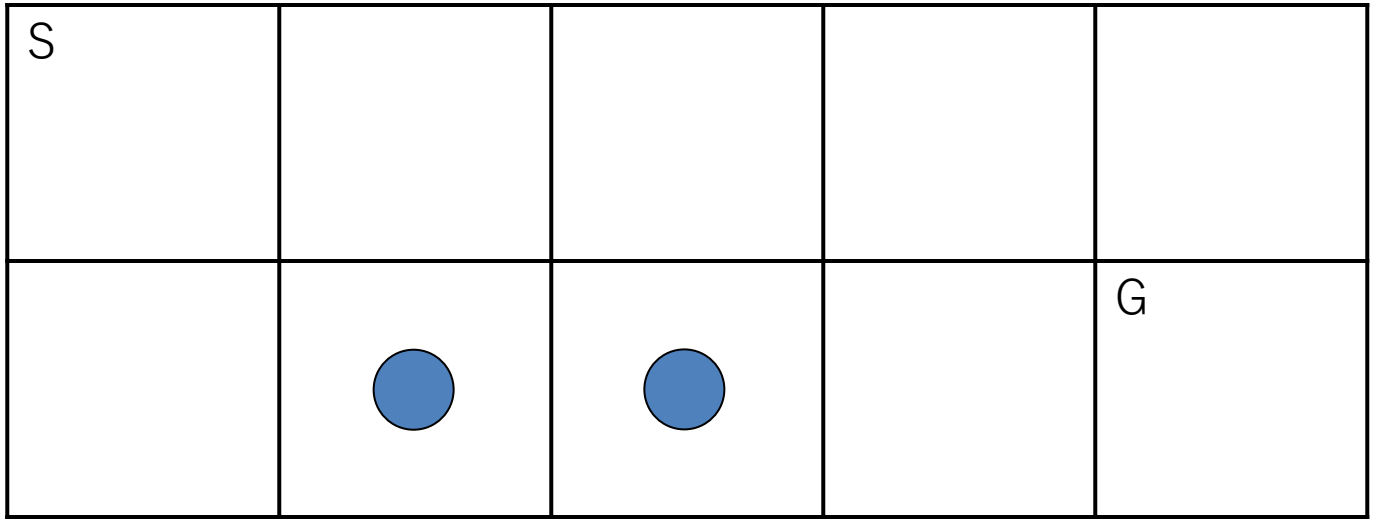


$$Q(6, right) \leftarrow 0 + 0.5[0 + 0.1 \times 0]$$

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1か0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

行動a：(up, down, left, right), 報酬r：(G：1, 他：0)

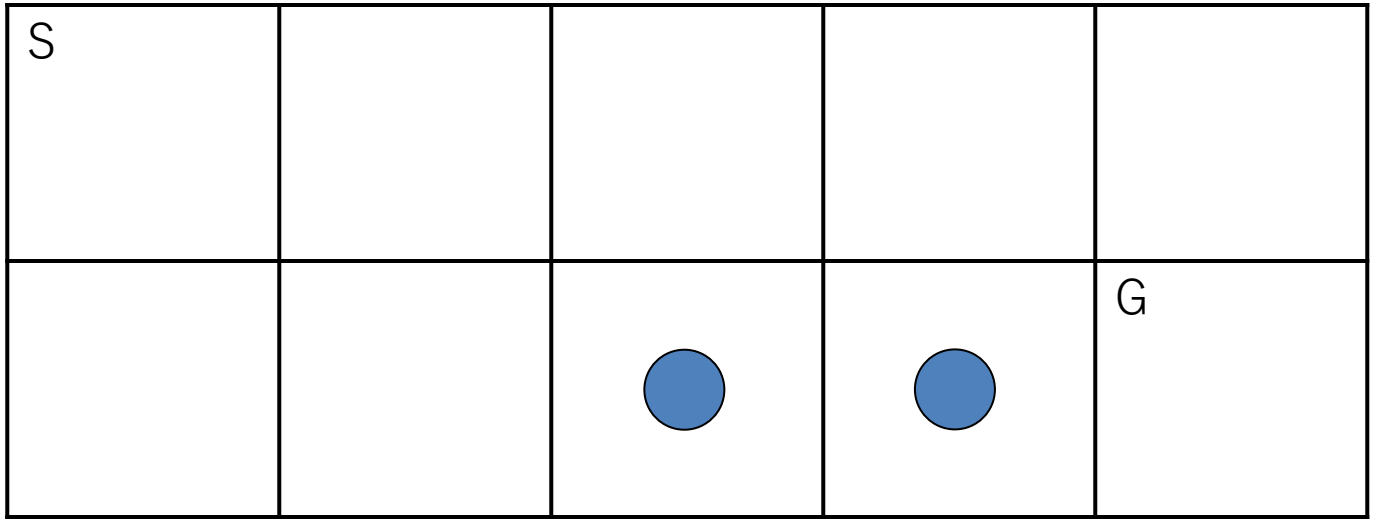


$$Q(7, right) \leftarrow 0 + 0.5[0 + 0.1 \times 0]$$

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1\text{か}0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

行動a：(up, down, left, right), 報酬r：(G：1, 他：0)

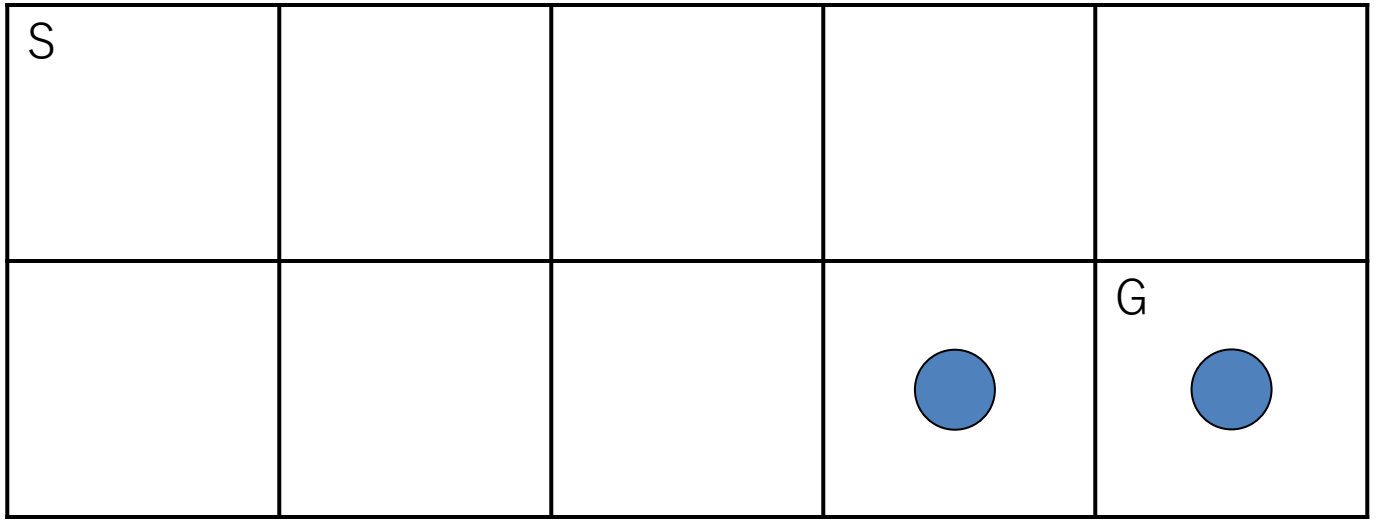


$$Q(8, right) \leftarrow 0 + 0.5[0 + 0.1 \times 0.5]$$

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1か0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

行動a：(up, down, left, right), 報酬r：(G：1, 他：0)



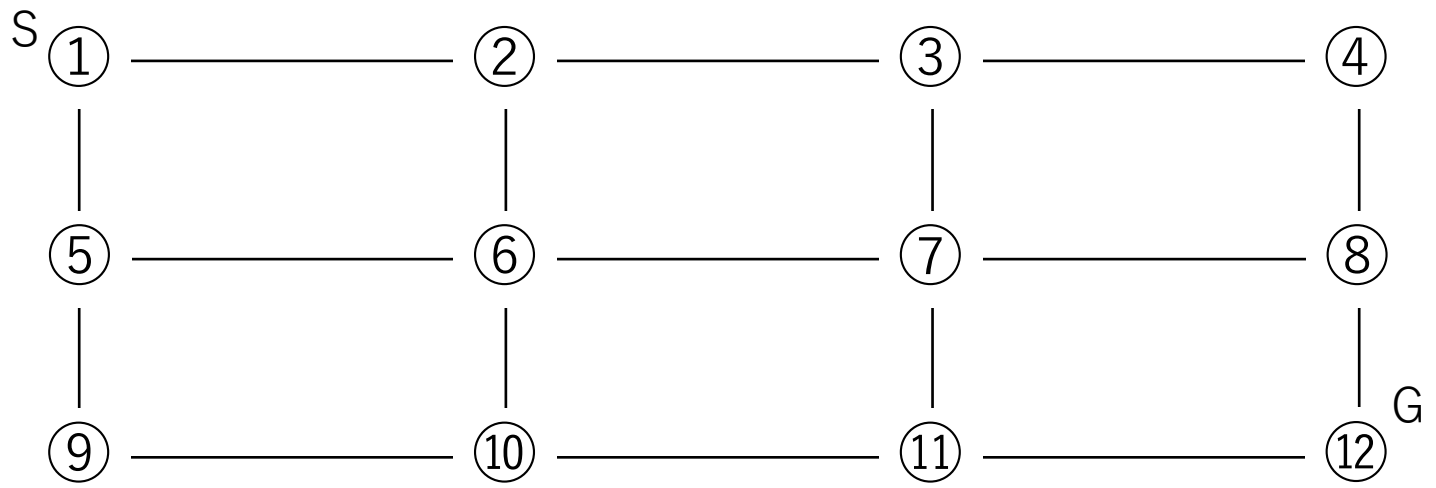
$$Q(9, right) \leftarrow 0.5 \times 0.5 + 0.5[1 + 0.1 \times 0]$$

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a')]$$

$$r = 1\text{か}0 \quad \alpha = 0.5 \quad \gamma = 0.1$$

練習問題14-1

- Q学習を用いて以下の経路図のスタート（①）地点からゴール地点（⑫）を目指す
- ①→②→③→⑦→⑧→⑫という状態遷移を3回繰り返した後の0でないQ値をすべて答えよ
- 報酬、Q値の更新式は前スライドのものと同じ



制御プログラミングの自動化・省力化

- タスク遂行のためのプログラミングを強化学習で
  自動化することにより、設計者の負担軽減が期待できる

ハンドコーディングよりも優れた解

- 試行錯誤を通じて学習するため、 人間の専門家が
  得た解よりも優れた解を発見する可能性がある

- 特に不確実性（摩擦やガタ、振動、誤差など）や
  計測が困難な未知パラメータが多い場合、
  人間の常識では対処し切れないことが予想される場合に有効

自律性と想定外の環境変化への対応

- 機械故障などの急激な変化や、プラントの経年変化のような緩慢な変化など、 予め事態を想定してプログラミングしておくことが困難な環境の変化に対しても自動的に追従

- 宇宙や海底など、通信が物理的に困難な場合に有効

まとめ

- 試行錯誤から学ぶ強化学習の概要を学んだ。
- 強化学習のモデル化としてのマルコフ決定過程について学んだ。
- Q-learningによるマルコフ決定過程における方策の学習アルゴリズムを学んだ。

復習問題

1. 強化学習の目的は何を探索する事か？

2. マルコフ決定過程の最適性は何によって評価されるか？

3. Ｑ学習におけるＱ値は何によって計算されるか？

4. 教師あり学習では教師信号が与えられるが強化学習では何が与えられるか？

次回の講義

- 定期試験に向けた復習
- 研究紹介