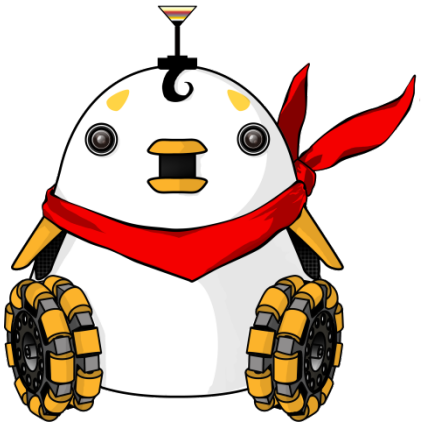
人工知能

第8回　復習問題と解説

立命館大学 情報理工学部

谷口彰



復習・補足説明

- 強化学習：Q学習
  - 演習８－４
  - 小テスト
  - 演習8-3* ベルマン方程式[証明]
  - 強化学習の分類とその発展*
- 確率とベイズ理論の基礎
  - 復習
  - 小テスト
- 復習問題（第８回 小テスト）

8.4.1 Q 学習

- 最適行動価値関数の確定遷移に対して

$$Q^*(s_t, a_t) = r_{t+1} + \gamma \max_{a_{t+1} \in A} Q^*(s_{t+1}, a_{t+1})$$

- 学習アルゴリズム
  - αは学習率

$$Q\ (s_t, a_t) \leftarrow Q\ (s_t, a_t) + \alpha \delta_t$$

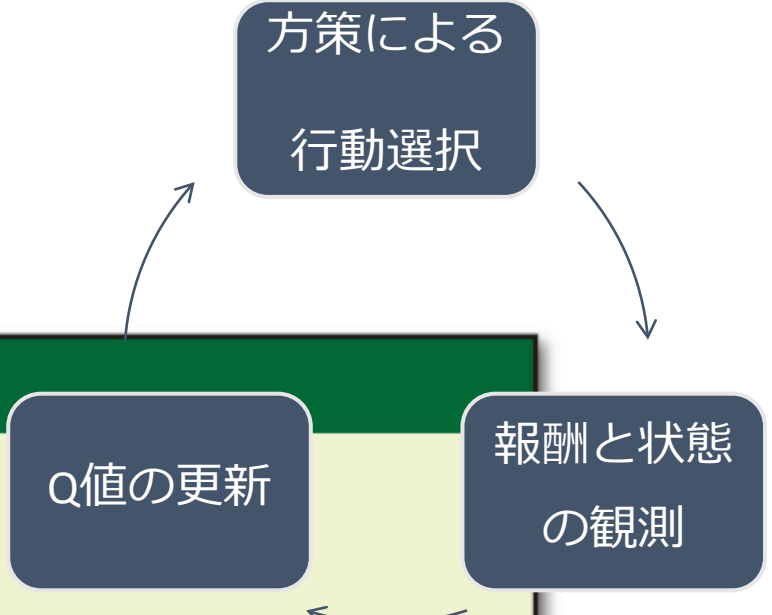
- TD誤差(Temporal difference error)

$$\delta_t = r_{t+1} + \gamma \max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$$

Q-learning　（これを繰り返してQ値を収束させる）

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

Algorithm



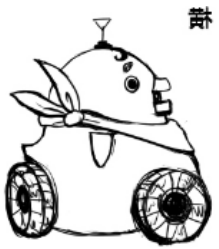
Algorithm 8.1 Q 学習

1. Q 値を初期化する.
2. **for** $i = 1$ to L **do**
3. 　　時刻 $t = 1$ として，s_0 を観測する.
4. 　　**repeat**
5. 　　方策 π に従って a_t を選択して行動する.
6. 　　環境から r_{t+1} と s_{t+1} を観測する.
7. 　　Q 学習の更新式 (8.20) に従って $Q(s_t, a_t)$ の値を更新する.
8. 　　時刻 $t \leftarrow t + 1$ とする.
9. 　　**until** ゴールに到達する，もしくは，終了条件に達する.
10. **end for**

演習8-4 Q学習の1-stepを追って見る.

select

行動a_t	Q値
右	8
左	10
停止	5



S_t → S_{t+1}

$r_{t+1}=4$

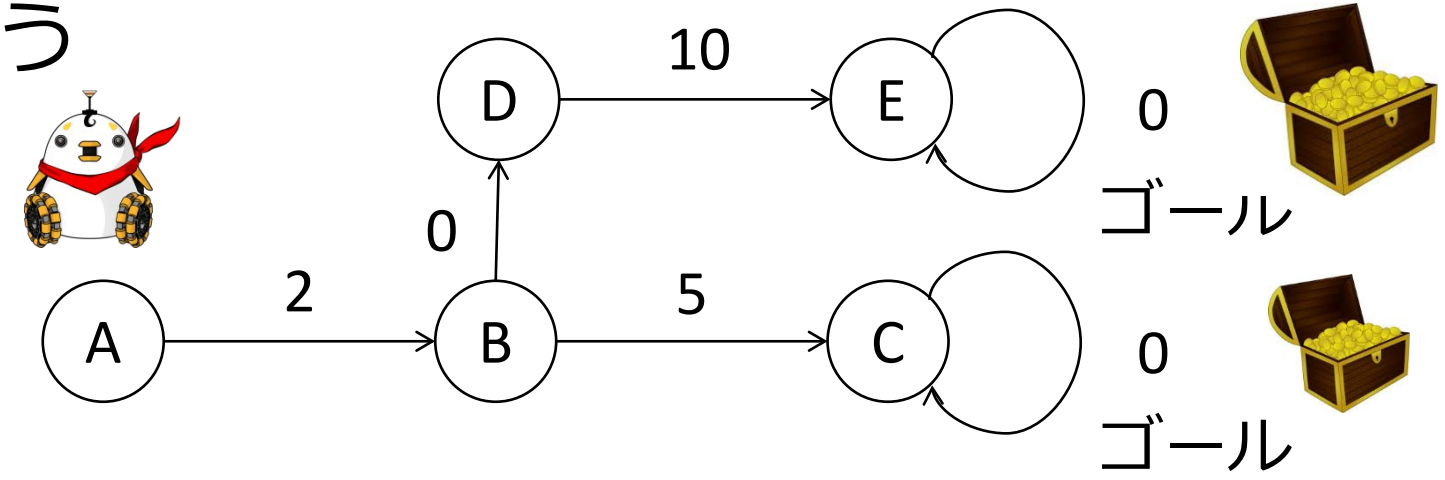
行動a_t	Q値
右	10
左	8
停止	5

ホイールダック２号は状態Stで行動「右」をとった結果St+1に遷移した.
それぞれの状態での現在の学習中の行動価値の値は表のとおりである.
割引率は0.9とする.

1. TD誤差δtはいくらか？
2. この1stepで表の内，どのQ値がどれだけ変わるか？学習率αを0.5として示せ.

第７回　小テスト

- 以下の分かれ道がある状態空間を考えよう



- 方策１　「右へ行けたら右，だめなら上」
- 方策２　「上へ行けたら上，だめなら右」

第７回　小テスト

- 問１
  前スライドの方策１と２におけるAからの状態遷移をそれぞれ以下のように表現せよ。

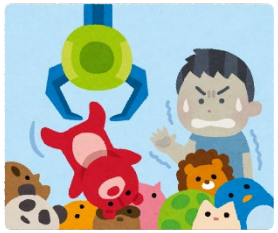
  A→B→…

- 問２
  方策１と２における割引累積報酬を$\gamma = 0.4$、$\gamma = 1.0$の場合について計算して以下の表にまとめ、その結果について考察せよ。

	A	B	C	D	E
方策1					
方策2					

演習8-3* ベルマン方程式[証明]

- 価値関数の定義式を用いて，下記のベルマン方程式が成立することを示せ.



ちょっと難しいですが考えてみてください.

$$V_\pi(s) = \sum_a \pi(s,a) \sum_{s'} P(s_{t+1} = s' | s_t = s, a_t = a)[r_{t+1} + \gamma V_\pi(s')]$$

価値関数の定義式

$$V_\pi(s) = E_\pi[R_t | s_t = s] = E_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s]$$

8.5.1-3 強化学習の分類

1. モデルフリーとモデルベース
   - P(s'|s,a)を学習するか否か？ Q学習はモデルフリー
2. 価値ベースと方策ベース
   - 方策を明示的に学習するか否か？ Q学習は価値ベース
3. On-policyとOff-policy
   - On-policyだとその方策で得た経験しかその方策の更新に使えない． Q学習はOff-policy

8.5.4 確率推論としての制御(control as inference)

将来に得られる累積報酬を最大化する行動を計画するという強化学習の問題が，将来にわたって最適な状態であり続けるとすれば，自らが今後とる行動はどのような行動系列$a_{t:T}$ であるかを**ベイズ推論**する問題へと変換される．

8.5.5 状態表現学習と深層強化学習

- マルコフ決定過程（MDP）を前提として多くの強化学習の手法は構築されるが，そもそも対象問題において何をMDP における状態s や行動a とするかは強化学習アルゴリズムを適用する以前に決定しなければならない問題
- 2013 年にムニらはDQN（Deep Q-Network）を提案して，Atari ゲームの画像情報をそのまま入力とした強化学習によってビデオゲームをプレイすることのできるゲームAI を作れることを示した.
- ディープラーニングと強化学習の融合（**深層強化学習**）

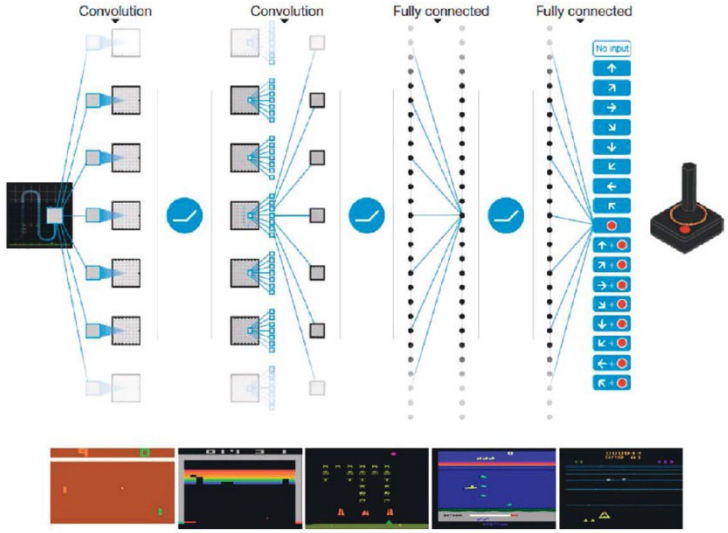


図 8.6 Atari ゲームをプレイする DQN ネットワーク構造

（上） V. Mnih et al.: Human-level control through deep reinforcement learning, *Nature*, 518(2015)529-533.

（下） V. Mnih et al.: Playing Atari with deep reinforcement learning, *NIPS Deep Learning Workshop*, 2013.

復習・補足説明

- 強化学習：Q学習
  - 演習８－４
  - 小テスト
  - 演習8-3* ベルマン方程式[証明]
  - 強化学習の分類とその発展*
- 確率とベイズ理論の基礎
  - 復習
  - 小テスト
- 復習問題（第８回 小テスト）

7.2.1 マルコフ過程：状態のみの確率システム
7.2.2 グラフィカルモデルとマルコフ性

■マルコフ性

$$P(s_{t+1}|s_{1:t}) = P(s_{t+1}|s_t) \qquad (6.25)$$

■マルコフ過程

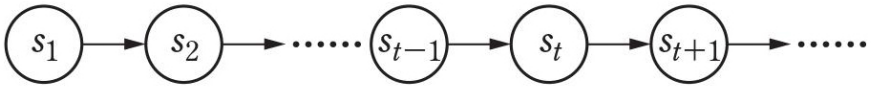


図 7.6　マルコフ過程のグラフィカルモデル

■マルコフ決定過程

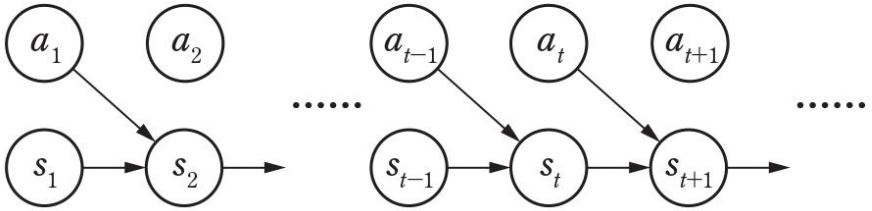


図 7.8　マルコフ決定過程のグラフィカルモデル

7.2.2 状態遷移確率

□状態遷移確率 (transition probability)

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,1} & p_{3,2} & p_{3,3} \end{pmatrix} = \begin{pmatrix} 0.1 & 1.0 & 0.5 \\ 0.5 & 0 & 0 \\ 0.4 & 0 & 0.5 \end{pmatrix} \tag{6.23}$$

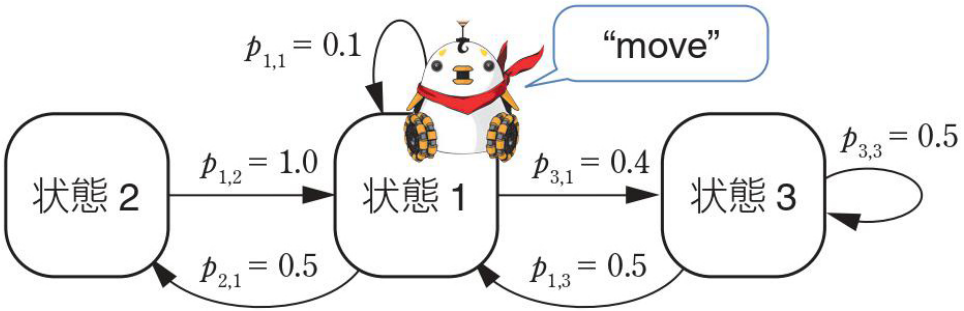
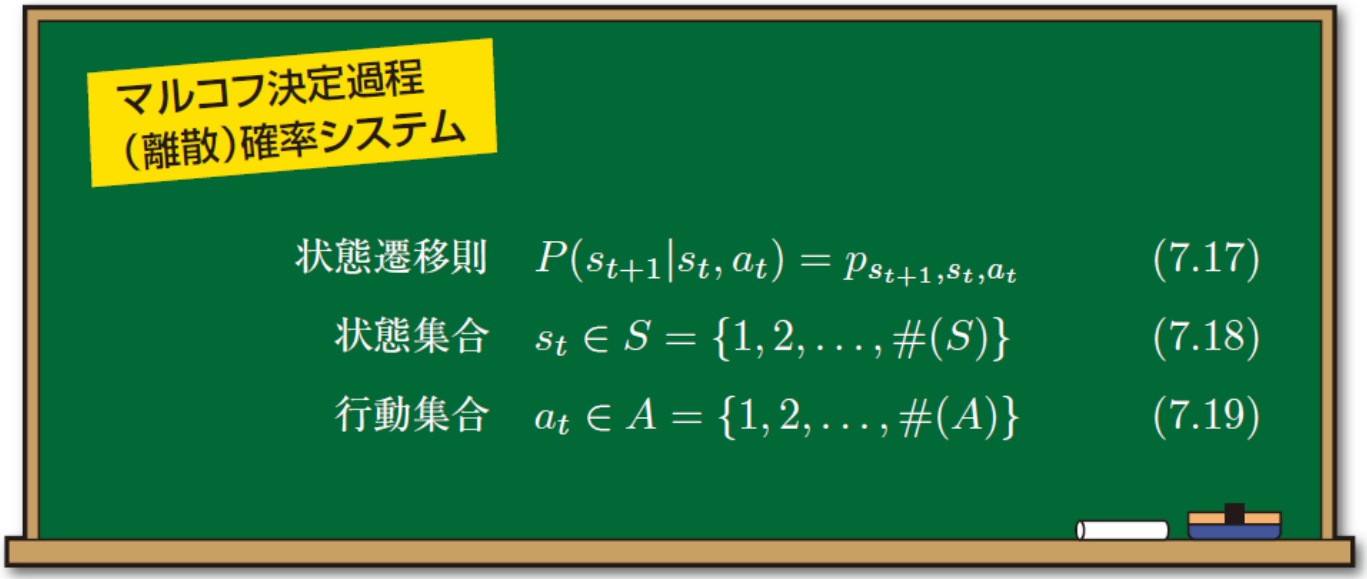


図 7.7 状態遷移確率を表すグラフ

7.2 確率システムの表現

- 次状態が現在の状態と行動に依存して確率的に決定するシステムのことを，**確率システム(stochastic system)**と呼ぶ.
- 確率システムの場合は状態遷移則が確率的になるため，**確率分布**による表現を用いる.



7.2.3 行動選択に依存した状態遷移確率

- 例えば行動として，A = {“stop”,“move”} の2 種類があり，a_t = “stop” の際にロボットは動かないとする.

$$\mathbf{P} = \left(\overbrace{\begin{pmatrix} p_{1,1,\text{move}} & p_{1,2,\text{move}} & p_{1,3,\text{move}} \\ p_{2,1,\text{move}} & p_{2,2,\text{move}} & p_{2,3,\text{move}} \\ p_{3,1,\text{move}} & p_{3,2,\text{move}} & p_{3,3,\text{move}} \end{pmatrix}}^{a_t=\text{“move”}}, \overbrace{\begin{pmatrix} p_{1,1,\text{stop}} & p_{1,2,\text{stop}} & p_{1,3,\text{stop}} \\ p_{2,1,\text{stop}} & p_{2,2,\text{stop}} & p_{2,3,\text{stop}} \\ p_{3,1,\text{stop}} & p_{3,2,\text{stop}} & p_{3,3,\text{stop}} \end{pmatrix}}^{a_t=\text{“stop”}} \right)$$

$$= \left(\begin{pmatrix} 0.1 & 1.0 & 0.5 \\ 0.5 & 0 & 0 \\ 0.4 & 0 & 0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) = (\mathbf{P}_{\text{move}}, \mathbf{P}_{\text{stop}}) \qquad (7.20)$$

ここで $p_{s_{t+1},s_t,a_t} = P(s_{t+1}|s_t, a_t)$ とする.

状態遷移の確率計算

$$P(s_{t+1}|a_t) = \sum_{s_t} P(s_{t+1}, s_t|a_t) = \sum_{s_t} P(s_{t+1}|s_t, a_t)P(s_t) \qquad (7.21)$$

第６回　小テスト

- 図7.7の状態遷移を前提としたときに，ホイールダック２号が初めに状態1に居たとして，move, stop, move という3つの行動を行った場合に，その後，ホイールダック２号が状態3にいる確率を求めよ.

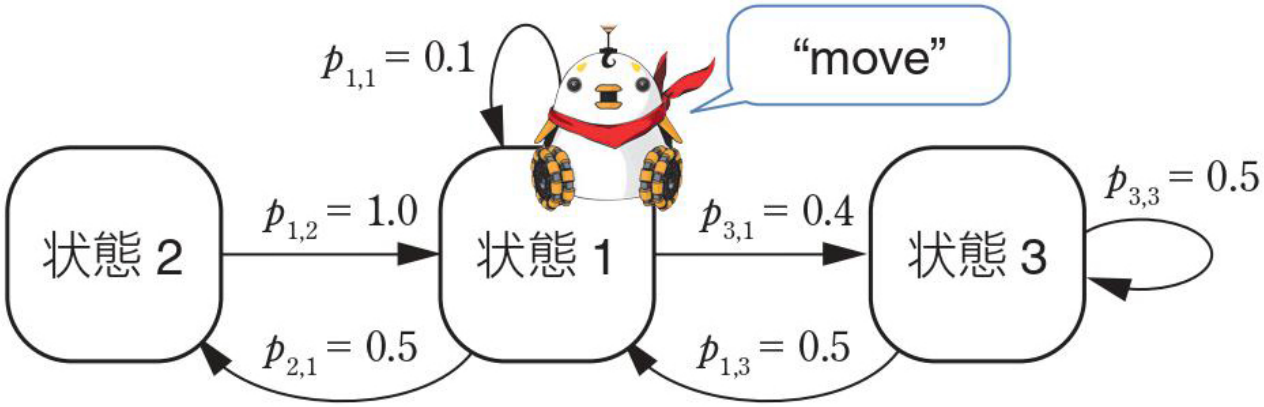


図 7.7　状態遷移確率を表すグラフ

復習・補足説明

- 強化学習：Q学習
  - 演習８－４
  - 小テスト
  - 演習8-3* ベルマン方程式[証明]
  - 強化学習の分類とその発展*
- 確率とベイズ理論の基礎
  - 復習
  - 小テスト
- 復習問題（第８回 小テスト）

復習問題（第8回 小テスト）

- 人工知能の基礎
- 状態空間と基本的な探索
- 最適経路の探索
- 割引累積報酬の計算

今回の小テストは、以下の復習問題を解いて提出してください。

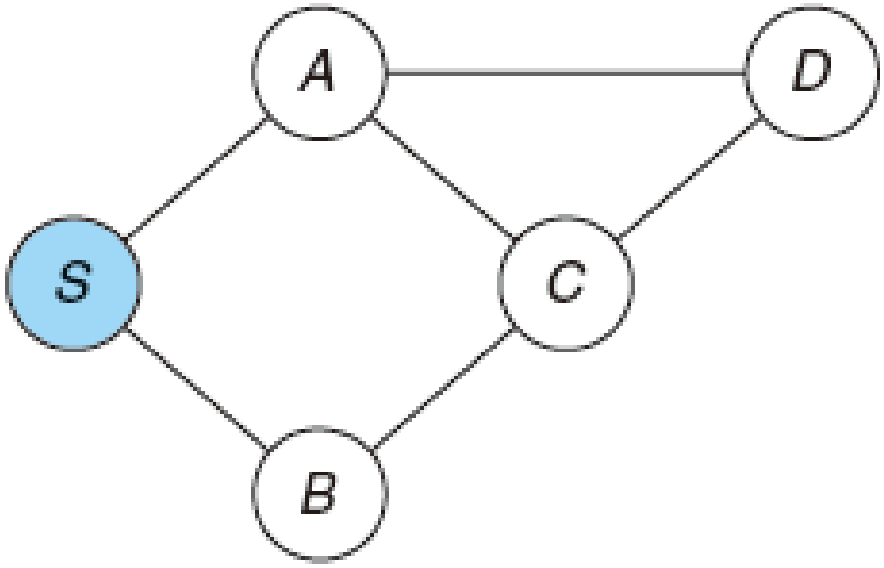
問1　人工知能の基礎

下記の問にそれぞれ答えよ.

1. チューリングテストとはなにか説明せよ.
2. フレーム問題とはなにかを説明せよ.
3. 記号接地問題とはなにかを説明せよ

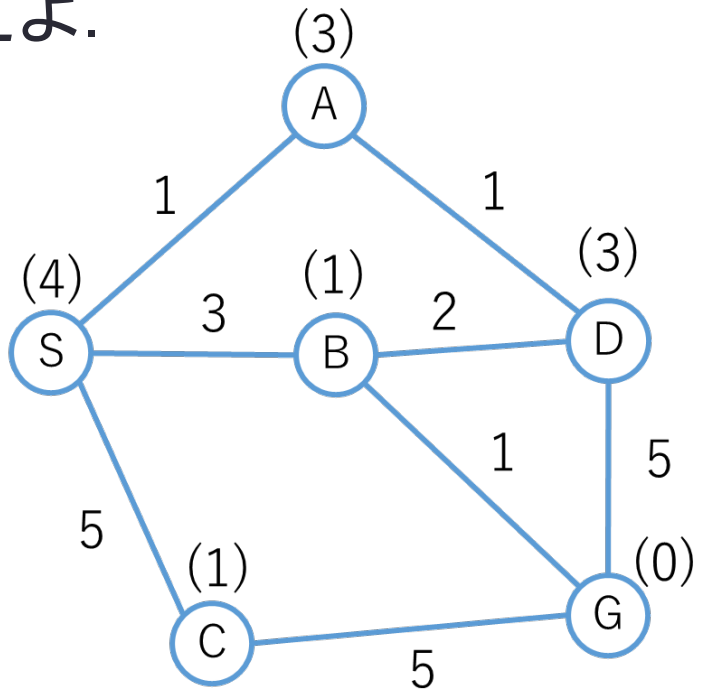
問2　状態空間と基本的な探索

- 図のグラフに関して，Sを初期状態として深さ優先探索と幅優先探索を行い，オープンリストとクローズリストの変化を示せ．このとき，探索においてアルファベットの並びの前を優先する．



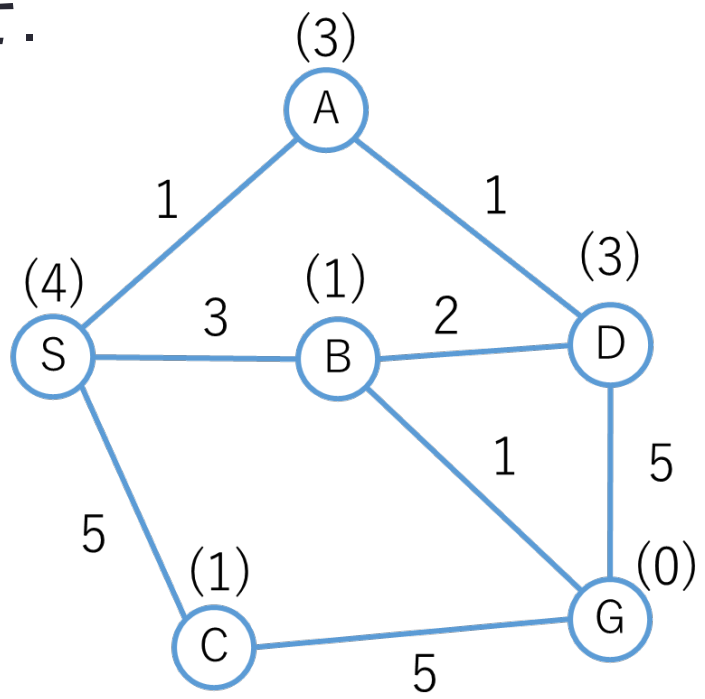
問3　最適経路の探索

- グラフについてノードSからノードGまでの最短経路を探索する．◯は各状態を表している．辺の横の数字はその辺を移動する際のコストを表し，ノードの上の数字は各ノードの予測評価値を表している．以下の（1）、（2）、（3）の問題に答えよ．



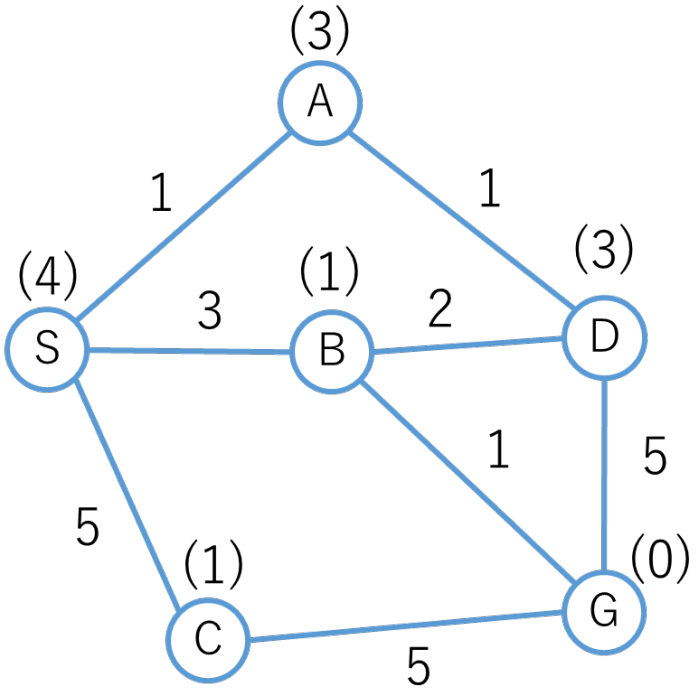
問3（1）

- A*アルゴリズムにより右のグラフを探索する．A*アルゴリズムを実行した際のオープンリストとクローズドリストの変化を逐次的に全て示せ．Sのみがオープンリストに入っている状態から始め，Gがクローズドリストに挿入される時点までを示せ．



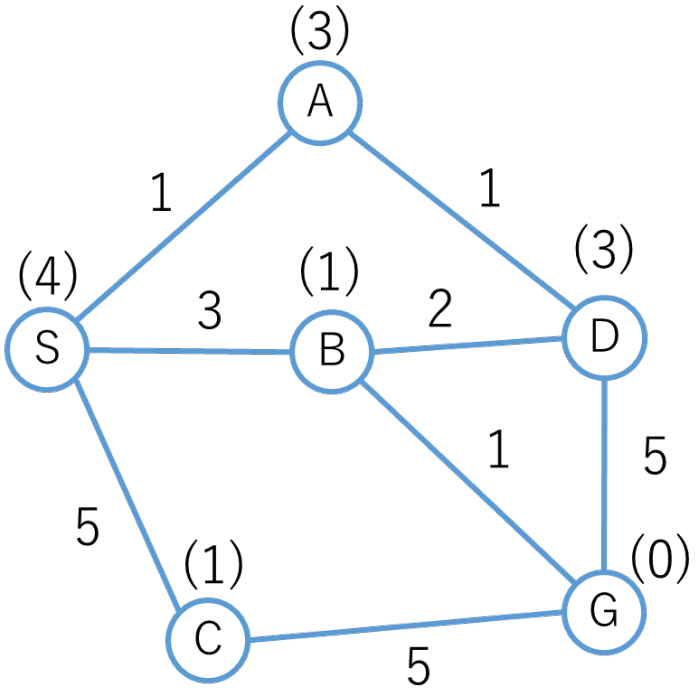
問3 （2）

- 右のグラフを最適探索で探索した時の解はA*アルゴリズムで探索した解と一致するか．一致するか否かと，そう考える理由を答えよ．

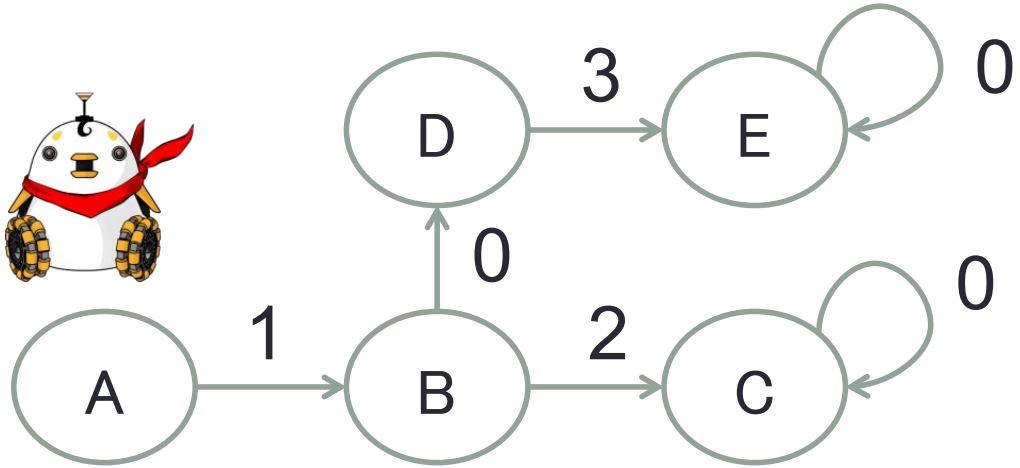


問3 （3）

- 最良優先探索によって右のグラフを探索した際の解を示せ.



問4



- 方策1は「右へ行けたら右，だめなら上」，
  方策2は「上へ行けたら上，だめなら右」という方策だとする．
  両方行けない場合はその場にとどまる．
- 割引率 $\gamma = 1$ の時のA,B,C,D,Eの状態における方策1に従う場合，方策2に従う場合，それぞれで割引累積報酬の値を求めよ．

	A	B	C	D	E
方策1					
方策2					

次回の講義

- 第9回 位置推定
  ベイズフィルタ
  - 8.1 位置推定の問題
  - 8.2 部分観測マルコフ決定過程
  - 8.3 ベイズフィルタ
  - 8.4 通路上のホイールダック2号の位置推定
    （ベイズフィルタ編）