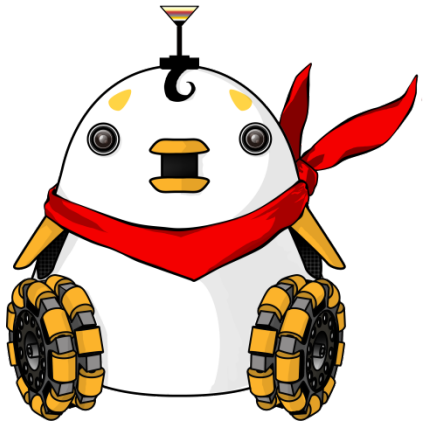
人工知能

第7回 多段決定(2)
強化学習

立命館大学 情報理工学部 知能情報学科
萩原良信



STORY 多段決定(2)

- 迷路に入る前に迷路の地図が完全にわかっているなどといった仮定はそもそもおかしいのではないだろうか．また，どの状態からどの状態の遷移でどれだけの利得が得られるという知識を事前に知っているという仮定も怪しいように思う．また，ある状態からある状態へ移動しようとするときに，その行動が必ず達成されるという仮定も疑わしい．場合によっては滑ることもあるだろう．では，何も利得や地図の知識を持たないままにホイールダック2号は経験のみに基づいて適切な経路を学習することはできるだろうか．

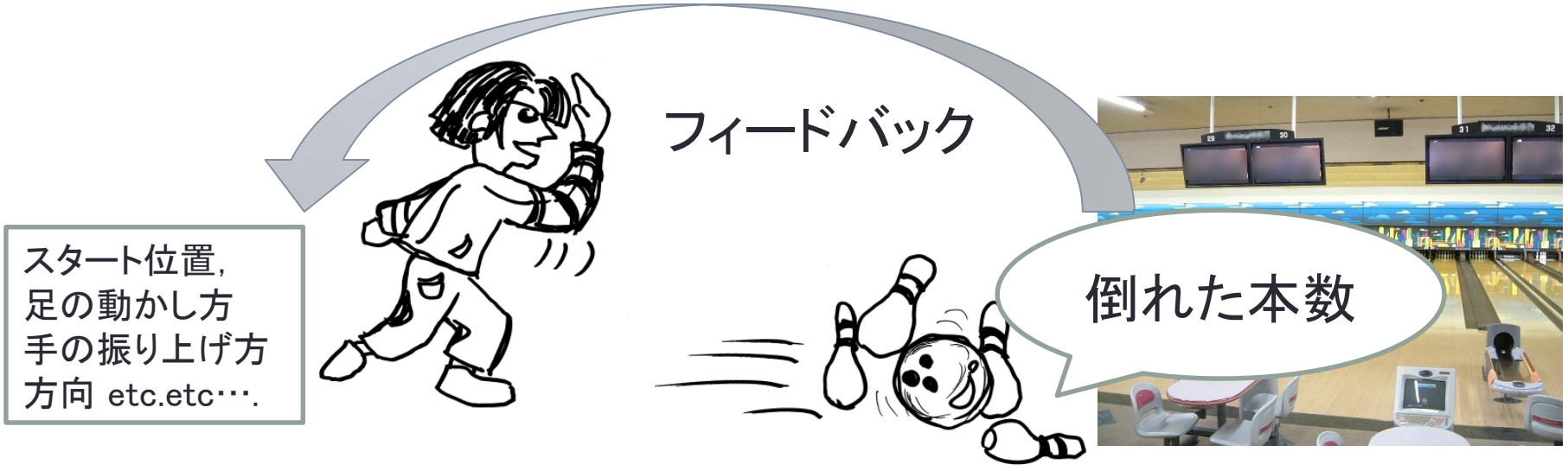
仮定多段決定(2)

- ホイールダック2号は迷路の完全な地図を持っていないものとする.
- ホイールダック2号は連続的な迷路の空間から適切な離散状態空間を構成できるものとする.
- ホイールダック2号は自分が状態空間のどの状態にいるかを認識できるものとする.
- ホイールダック2号は物理的につながっている場所・状態へは行動に応じて確率的に遷移するとする.

Contents

7.1.1 試行錯誤の中での学習

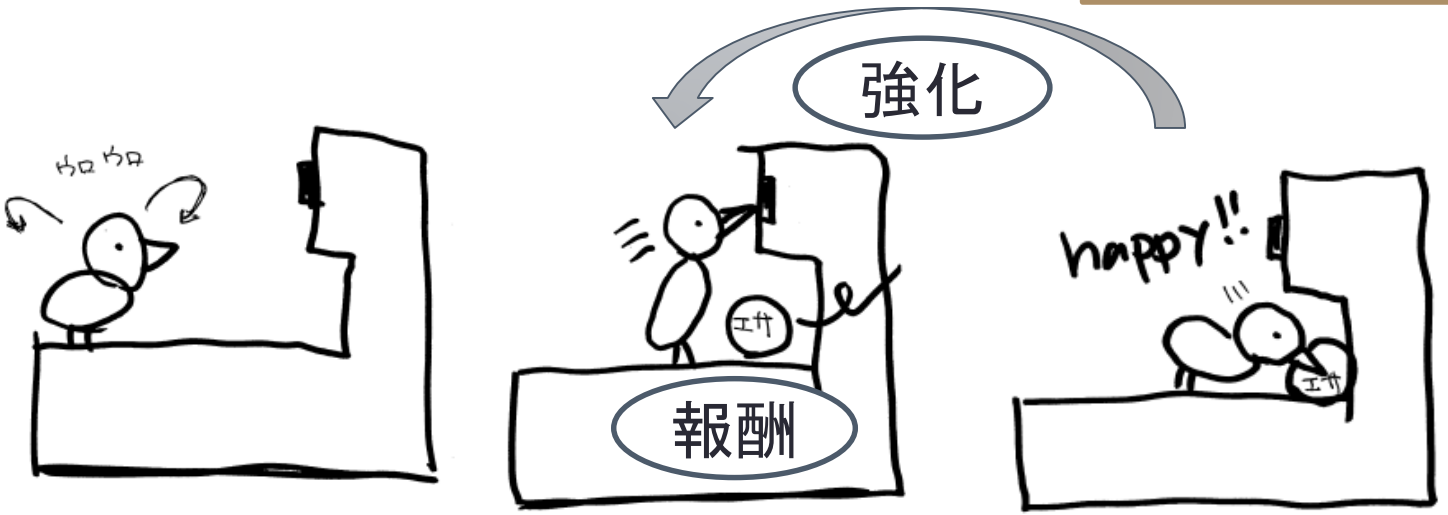
- 試行錯誤で学ぶ人間
  - 人間の様々な学習の進め方の中で，試行錯誤を通した学習がある．
  - やってみては，その結果・評価を観察し，徐々に「やり方」を改善していく．
  - 例）サッカーのフリーキック，ボーリング etc.etc.



7.1.1 オペラント条件づけ

- 自発的な試行錯誤の結果として得られる報酬によって行動形成がなされることを心理学でオペラント条件づけと呼ぶ.
- スキナー箱(Skinner 1938)

強化：その行動をとりやすくなること.



ハトはスイッチを押して餌を食べることを学習していく

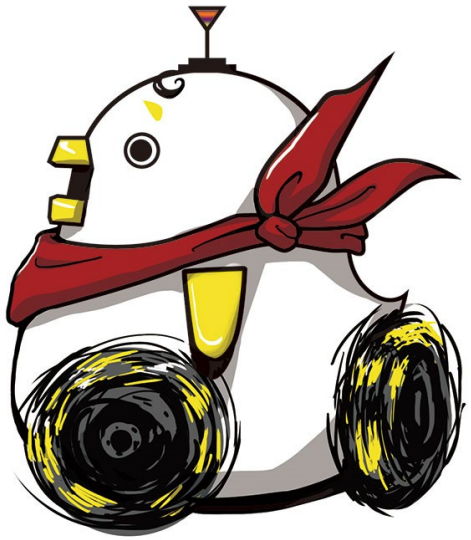
7.1.2 強化学習理論

- 試行錯誤による学習をロボットにさせるための機械学習法
- 強化学習は学習という語が含まれているが，動的計画法や制御理論における最適制御論などと近接した概念．
- 前回の動的計画法との相違点
  - はじめから状態空間や遷移則を与えないために，知識や環境の不確実性を扱わねばならず，そのために確定システムではなく確率システムとしてシステムをモデル化している．
  - 情報を得ながらの学習，つまり，オンラインでの学習を仮定している．

7.1.3 方策と価値

- 方策(policy)
  - ある状態にいたときに，どういう行動をどれほどの確率で選択するか．
- 価値関数(value function)
  - 状態や行動の価値

A*アルゴリズムや動的計画法と異なり，「経路」を求めることが問題ではなく，方策/価値関数を求めることが目的となる．

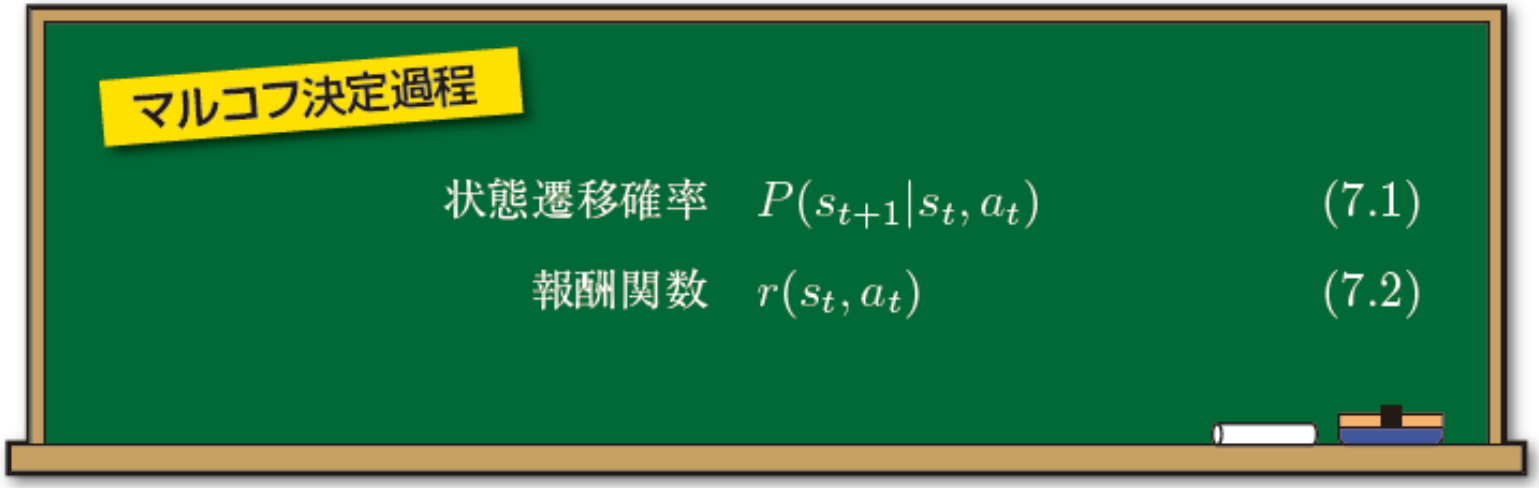


Contents

- 7.1 強化学習とは何か？
- 7.2 マルコフ決定過程
- 7.3 割引累積報酬
- 7.4 価値関数
- 7.5 学習方法の例：Q学習

7.2.1 状態遷移確率と報酬関数

- 強化学習はマルコフ決定過程(MDP，Markov Decision Process) に基づいて定式化される.



- 方策(policy)
  - 将来にわたって得られる報酬の期待値を最大化する方策を見つけることが強化学習の問題

$$\pi(s, a) = P(a_t = a|s_t = s)$$

Contents

- 7.1 強化学習とは何か？
- 7.2 マルコフ決定過程
- 7.3 割引累積報酬
- 7.4 価値関数
- 7.5 学習方法の例：Q学習

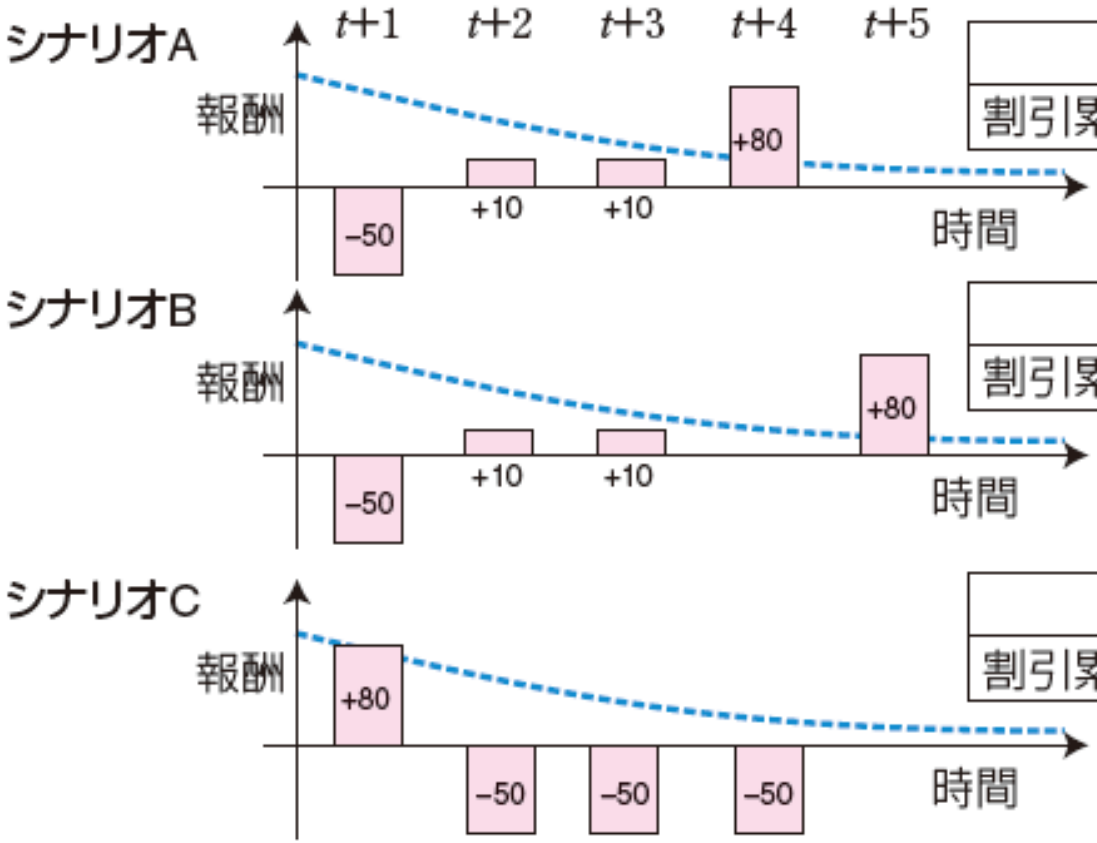
7.3.1 割引累積報酬の意味

- 割引累積報酬(discounted return) R_t

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

- γ ($0 \leq \gamma < 1$) は割引率(discount rate) と呼ばれる定数である.
- 割引累積報酬は基本的には将来にわたって得られる報酬の和になっているが，遠い未来であればあるほど，割り引いて換算される.
- $\gamma = 1$ では T→∞で発散する.
- r : 報酬

7.3.2 割引率と未来の報酬価値



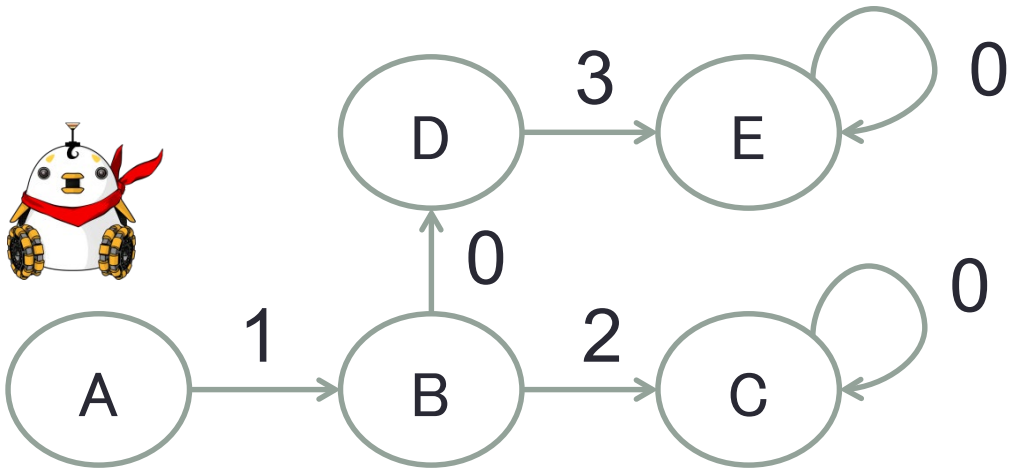
割引率：γ	1.0	0.9	0.5
割引累積報酬：R_t	50	25	−33

割引率：γ	1.0	0.9	0.5
割引累積報酬：R_t	50	14	−40

割引率：γ	1.0	0.9	0.5
割引累積報酬：R_t	−70	−42	36

図 7.3　割引累積報酬の割引率による違い

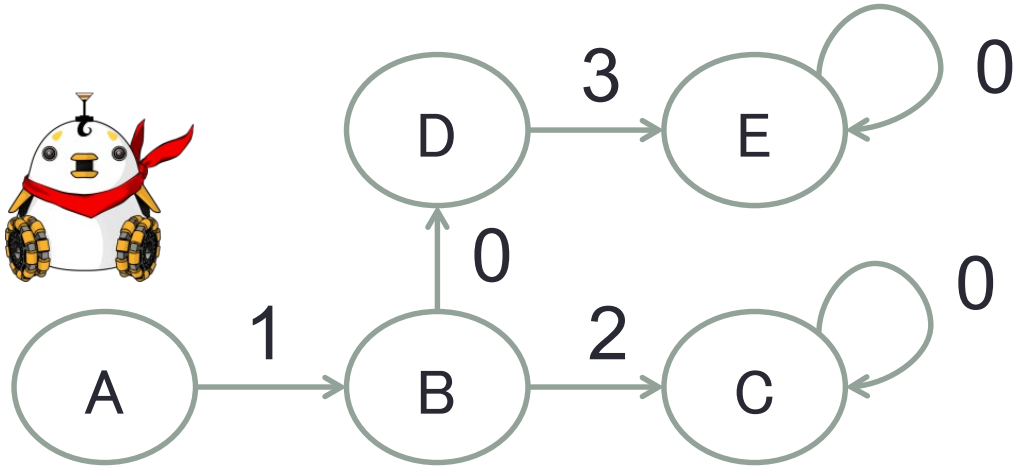
演習7-1割引累積報酬の計算



- 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 $\gamma = 0.5$ 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					

演習7-2 割引累積報酬の計算

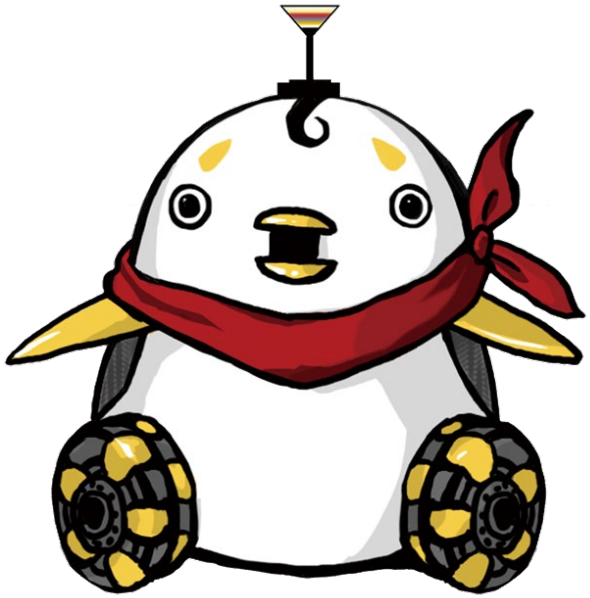


- 方策1は「右へ行けたら右，だめなら上」，
  方策2は「上へ行けたら上，だめなら右」という方策だとする．
  両方行けない場合はその場にとどまる．
- 割引率 $\gamma = 1$ の時のA,B,C,D,Eの状態における方策1に従う場合，方策2に従う場合，それぞれで割引累積報酬の値を求めよ．

	A	B	C	D	E
方策1					
方策2					

7.3.5 まとめ：割引率と報酬と評価値

- 割引率γ が異なれば，よりよい方策は異なる．
- 各状態における割引累積報酬は方策によって異なる．
- 割引累積報酬を方策の評価値と考えた場合には，その評価値は状態によって異なる．



Contents

- 7.1 強化学習とは何か？
- 7.2 マルコフ決定過程
- 7.3 割引累積報酬
- 7.4 価値関数
- 7.5 学習方法の例：Q学習

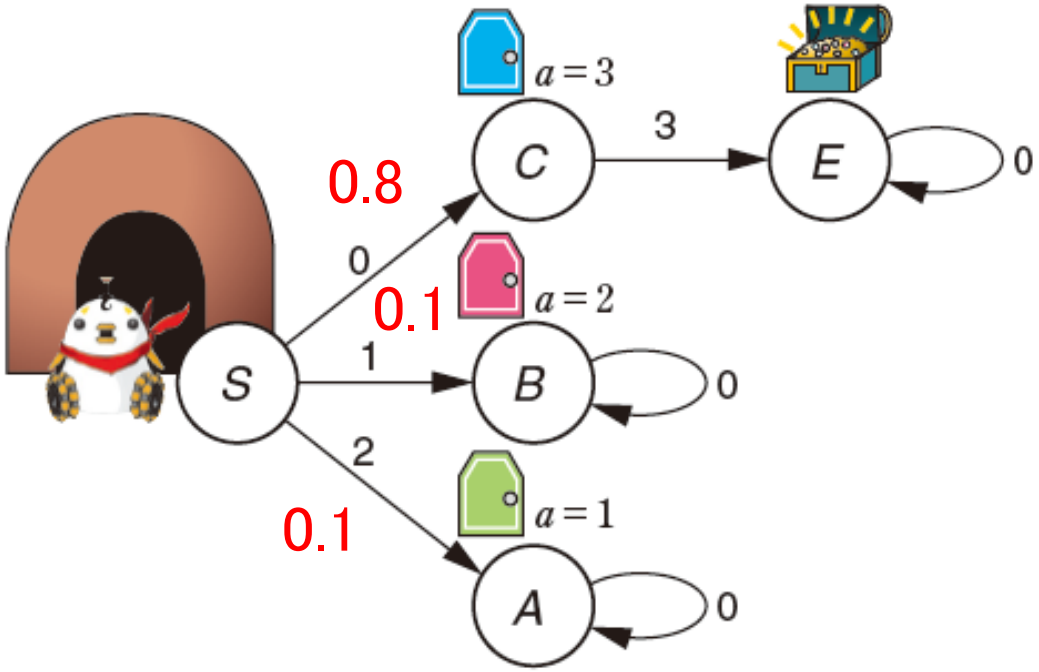
7.4.1 状態価値関数

- よりよい方策を学習するためには，正しく状態と行動の価値を見積もる必要がある．このために価値関数が定義される．

- 状態価値関数（state-value function）$V_\pi(s)$
  - 「その方策 π に従えば，その状態 s からスタートして将来にどれだけの割引累積報酬を得られるか」
  - E：期待値，R:割引累積報酬

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

7.4.2 ホイールダック2 号と分かれ道（確率編）



方策 π を行動の選択確率で表す
S→A : 0.1
S→B : 0.1
S→C : 0.8
割引率
γ =0.9

図 7.5 ホイールダック 2 号と分かれ道の先にある報酬

$$V_\pi(s = S) = 0.8 \times 2.7 + 0.1 \times 1.0 + 0.1 \times 2.0 = 2.46$$

状態価値関数の値を高める方策 π がより良い方策といえる

7.4.3 行動価値関数

- 行動価値関数(action-value function)

$$Q_\pi(s,a) = E_\pi[R_t|s_t = s, a_t = a] = E_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s, a_t = a] \tag{7.7}$$

- 状態価値関数と行動価値関数の関係式

$$V_\pi(s) = \sum_a \pi(s,a) Q_\pi(s,a)$$

- 最適行動価値関数

$$Q^*(s,a) \equiv Q_{\pi^*}(s,a) = \max_\pi Q_\pi(s,a) \tag{7.9}$$

未来はドンドン分岐する

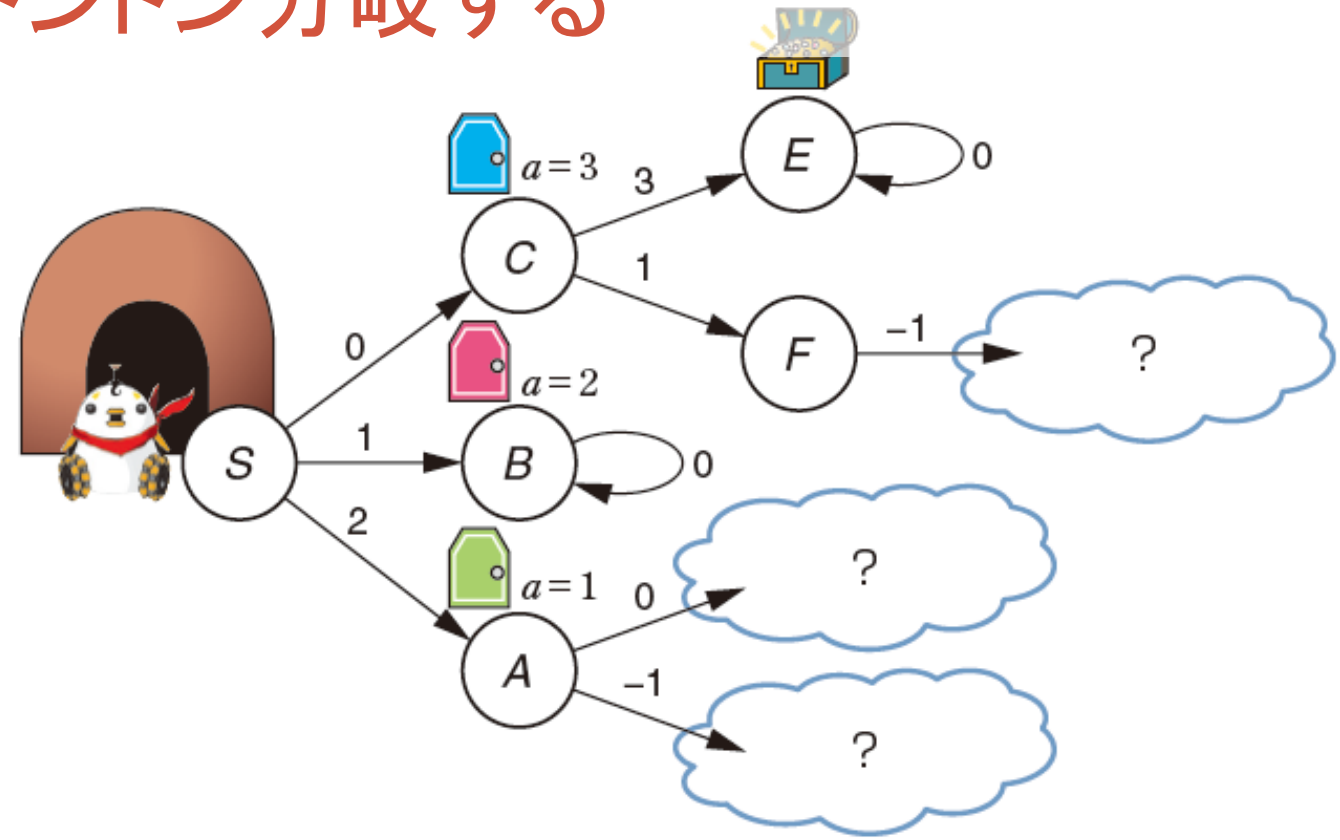
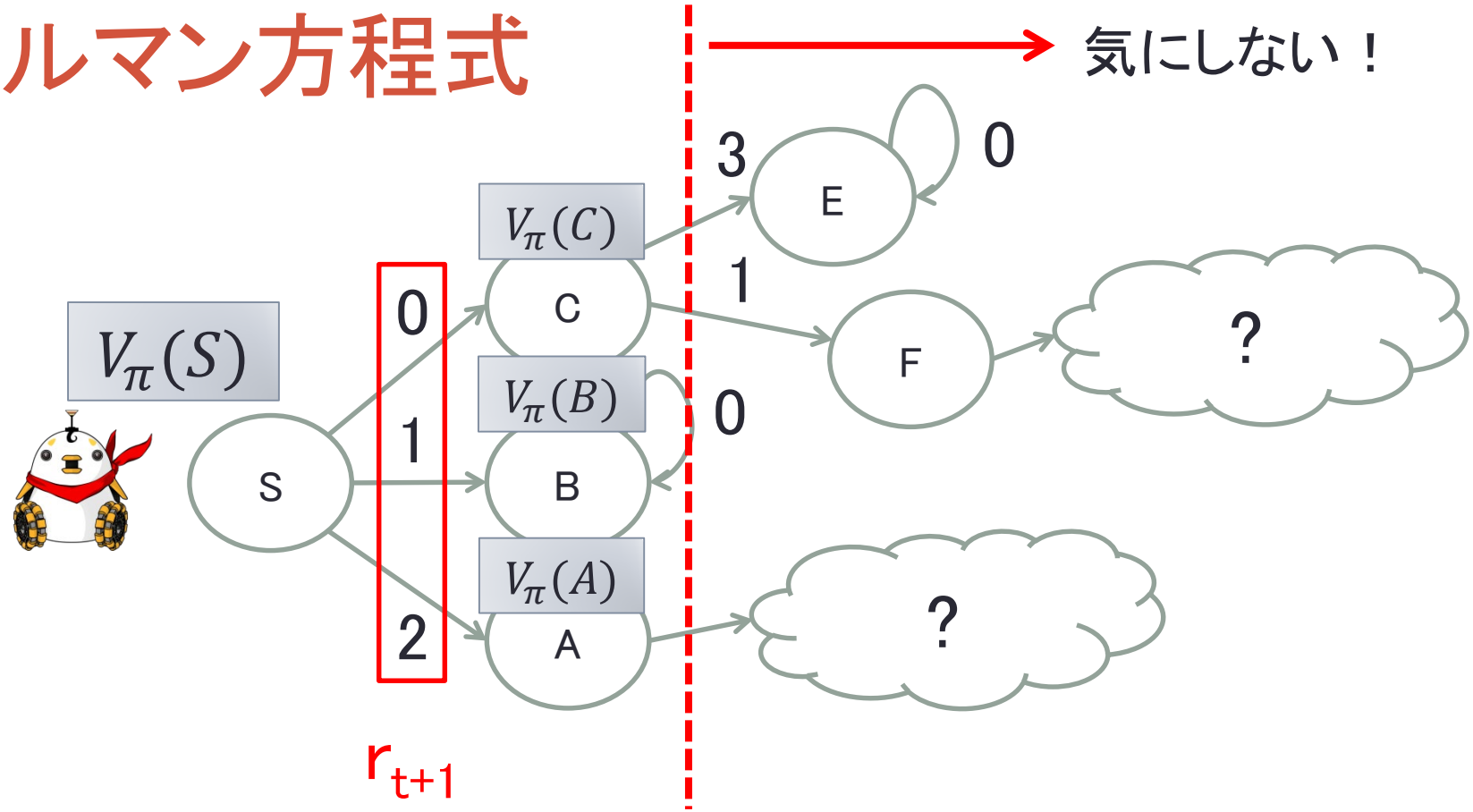


図 7.6 ホイールダック 2 号の未来はどんどん分岐する

- 視点
  1. 問題を簡単にする上で状態価値の間に良い性質は無いか？
  2. オンライン学習に変更するためのよい近似方法は無いか？

ベルマン方程式



- 現状態の状態価値は次の報酬と次状態の価値だけで定義出来る．下の式をベルマン方程式と呼ぶ．

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

行動価値関数のベルマン方程式

$$Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} V_\pi(s') P(s'|s,a) \tag{7.11}$$

$$V_\pi(s') = \sum_{a'} \pi(s',a') Q_\pi(s',a') \tag{7.12}$$

- ベルマン方程式に基づいて強化学習の問題を解く様々な手法が提案されている.
  - 例） SARSA，アクタークリティック法，Q学習など

Contents

- 7.1 強化学習とは何か？
- 7.2 マルコフ決定過程
- 7.3 割引累積報酬
- 7.4 価値関数
- 7.5 学習方法の例：Q学習

7.5.1 Q 学習

- 最適行動価値関数の確率遷移に対して

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

- 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

方策による行動選択

報酬と状態の観測

Q値の更新

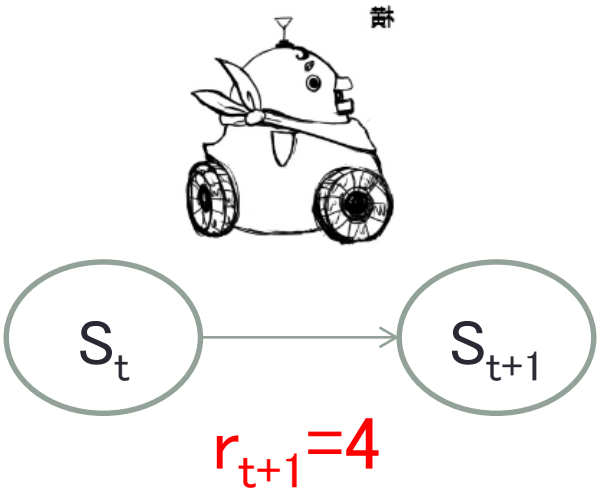
Algorithm 7.1 Q 学習

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7-4 Q学習の1-step

select

行動a_t	Q値
右	8
左	10
停止	5



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 δ_t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 多段階決定(2)

- 割引累積報酬と，その割引率の変化による影響について具体的な比較を通して学んだ.
- 割引累積報酬の期待値を表現する関数として状態価値関数と行動価値関数について学んだ.
- ベルマン方程式として適切な価値関数が満たすべき漸化式を得た.
- Q 学習のアルゴリズムとQ 学習における方策の決定方法について学んだ.

次回の講義

- 中間テスト
  - 範囲：今日までの内容
  - 教室：B201
  - 日時：５月５日　８：００－９：４０
  - 方式：選択問題と記述問題
  - アドバイス：講義内の演習や小テストは解けるようにしておこう。