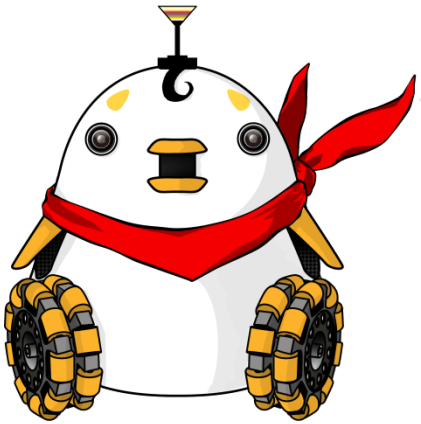
人工知能

第7回　計画と決定(2)
強化学習

立命館大学 情報理工学部
谷口彰



STORY　強化学習

- 迷路の3階にたどり着いたホイールダック２号は困っていた．地図を持っていないし，ゴールの位置もわからない．
- 迷路に入る前に迷路の**地図が完全にわかっているという仮定はそもそもおかしい**．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**．
- 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**．

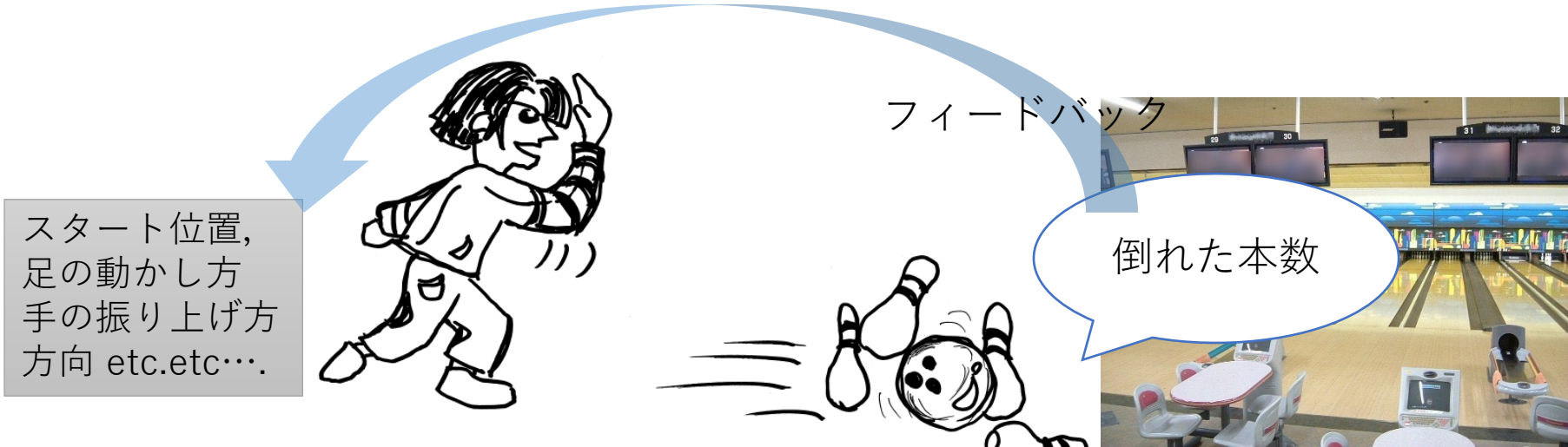
仮定　強化学習

- 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

8.1.1 試行錯誤の中での学習

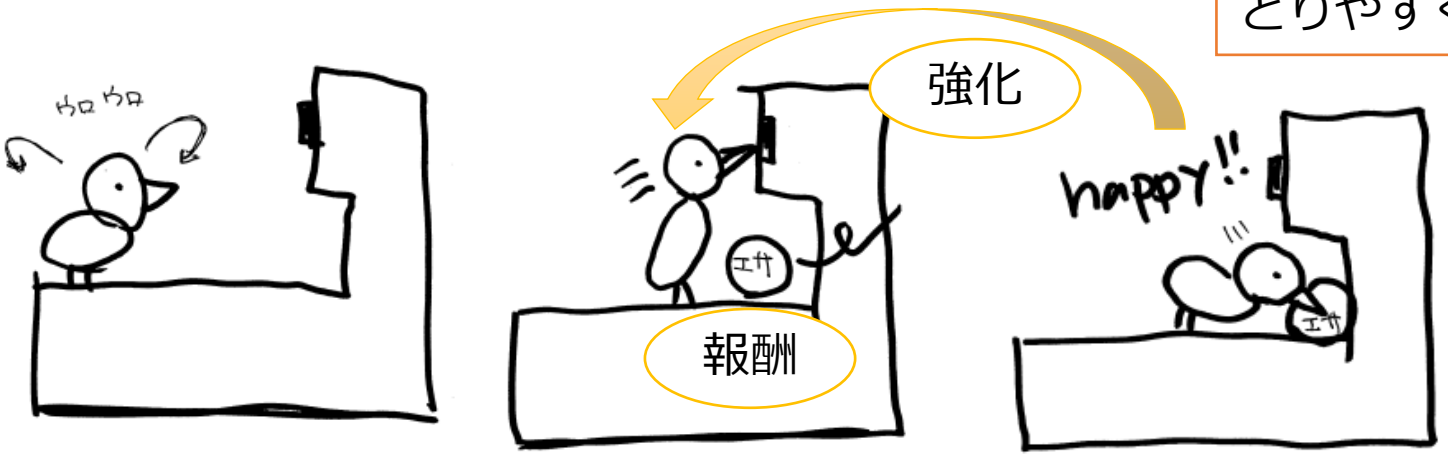
- 試行錯誤で学ぶ人間
  - 人間の様々な学習の進め方の中で，試行錯誤を通した学習がある．
  - やってみては，その結果・評価を観察し，徐々に「やり方」を改善していく．
  - 例）サッカーのフリーキック，ボーリング etc.etc.



8.1.1 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

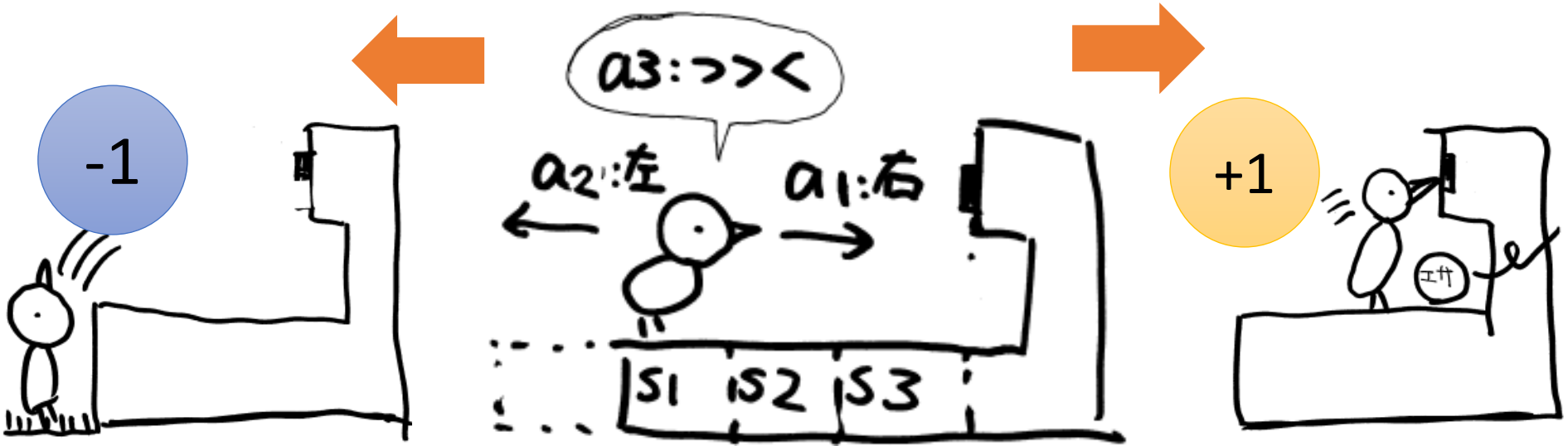
8.1.1 強化学習理論

- 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ので，知識や環境の不確実性を扱わなければならない.
  - 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を仮定している.

直感的な導入！（厳密性は気にせずに・・）

8.1.2 例：ハト型ロボットの学習

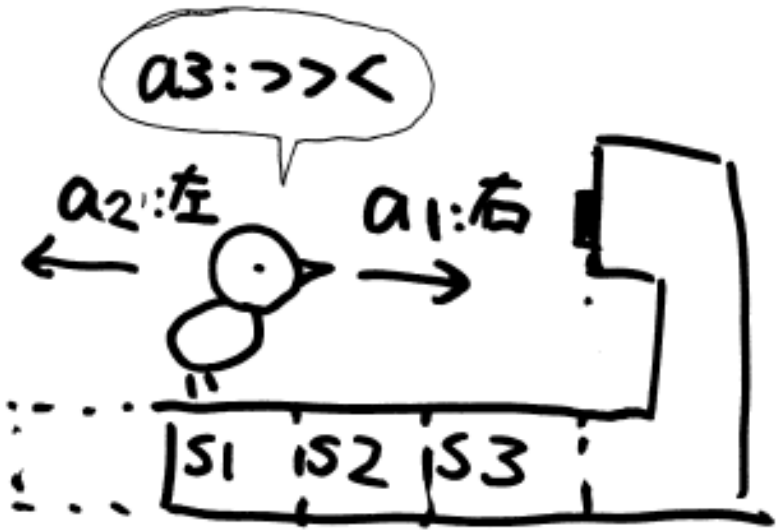
- 三つの状態S1,S2,S3があるものとし，それぞれに於いて３種類の行動a1:右, a2:左，a3:つつき により移動，動作するロボットを考える.
- ロボットは「どの状態ではどの行動をとるべきか？」ということを学ぶ.



状態と行動に対応した「価値」表

行動価値 Q(s1,a1)

	S1	S2	S3
a1: 右に行く	0	0	0
a2: 左に行く	0	0	0
a3: つつく	0	0	0



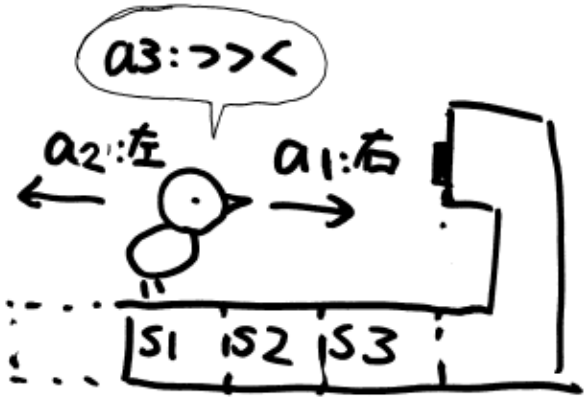
- それぞれの状態s1～s3で行動a1～a3をとることがどれだけ「うれしい」かを表わす「価値」の表を準備する.
- 各状態でどの行動をとれば良いか一目瞭然でわかるようになる.

まず，**報酬がもらえた時にもらえた報酬の値を，その状態での行動価値とする**というルールで「価値」表を書き換えてみる！！

[試行1]
s1 -> a1(右), s2 -> a2(左)・・・ただブラブラする.

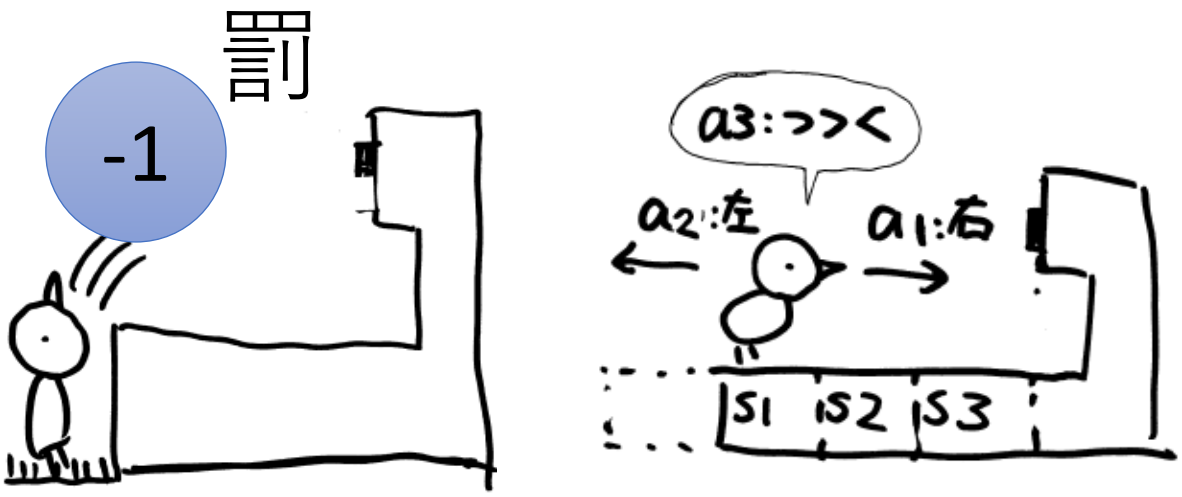
	S1	S2	S3
a1: 右に行く	0	0	0
a2: 左に行く	0	0	0
a3: つつく	0	0	0

報酬も罰ももらえないので
なにも変わらない



[試行2]
s1 -> a2(左)・・・落ちて罰を受ける.

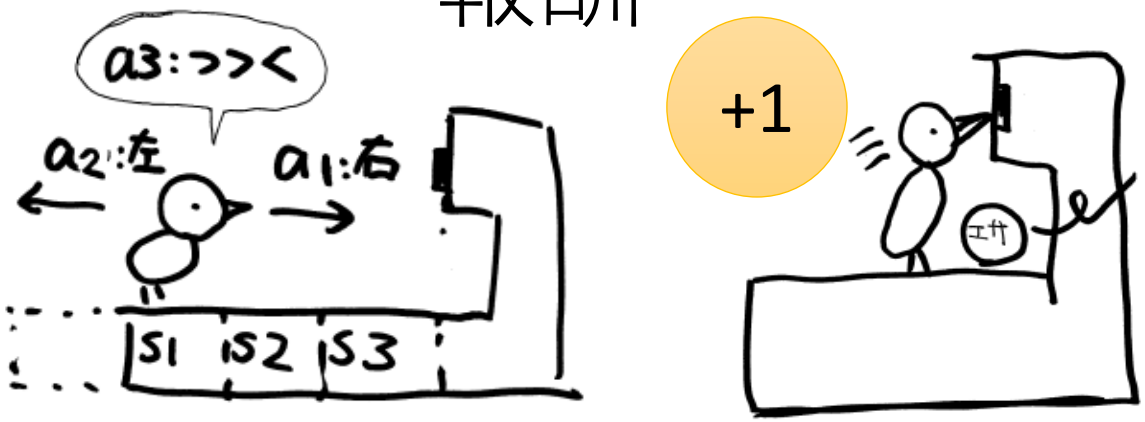
	S1	S2	S3
a1: 右に行く	0	0	0
a2: 左に行く	-1	0	0
a3: つつく	0	0	0



[試行3]
s1 -> a1(右) , s2-> a1(右), s3 -> a3(つつく)・・・
スイッチを押して餌を得る.

	S1	S2	S3
a1: 右に行く	0	0	0
a2: 左に行く	-1	0	0
a3: つつく	0	0	**+1**

報酬



即時報酬に基づく学習の落し穴

	S1	S2	S3
a1: 右に行く	0	0	0
a2: 左に行く	-1	0	0
a3: つつく	0	0	**+1**

- S2からの行動では絶対に報酬が得られないので，S2における行動の得点が変わらない．
- その結果， S2 でどのような行動をとれば良いのかは永遠に分からない．

・・・どうすればいいだろうか？

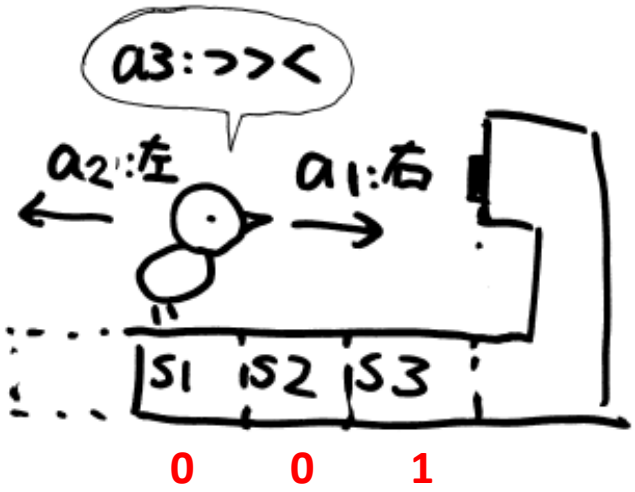
「状態価値」の導入

行動価値 $Q(s_1,a_1)$

	S1	S2	S3
a1: 右に行く	0	0	0
a2: 左に行く	-1	0	0
a3: つつく	0	0	+1
状態価値	0	0	**+1**

状態価値 $V(s_1)$

- **状態価値**
  - 「その状態がどのくらいうれしいか？」
  - 「その状態に来れば次にいくらの報酬を得られるか？」
- 価値のある状態に辿り着けばその半分を**割引きした価値**がその行動にはあるとして更新する.

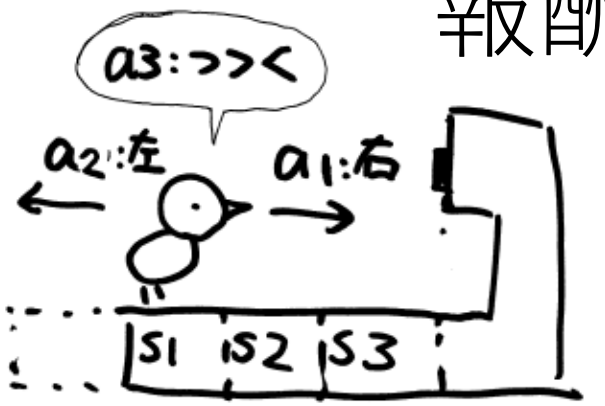


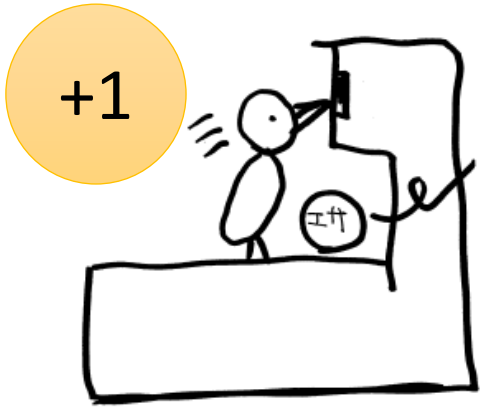
s1 -> a1(右) -> a1(右) -> a3(つつく)
・・・スイッチを押して餌を得る

	S1	S2	S3
a1: 右に行く	**+1/4**	**+1/2**	0
a2: 左に行く	-1	0	0
a3: つつく	0	0	+1
状態価値	+1/4	+1/2	**+1**

割引き

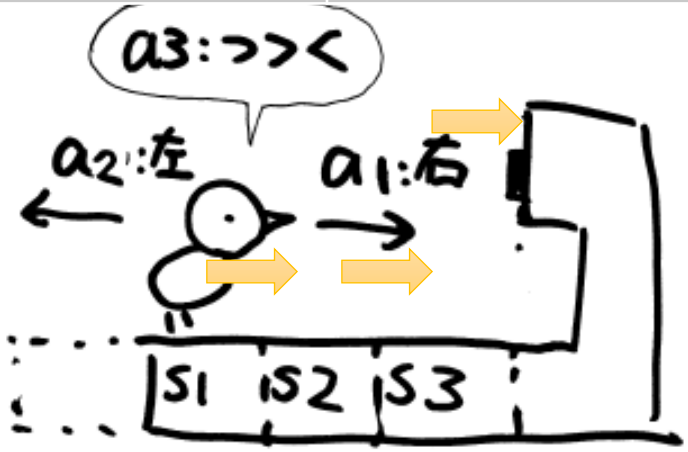
報酬





ロボットが各状態で最も価値があると判断した行動

	S1	S2	S3
a1: 右に行く	**+1/4**	**+1/2**	0
a2: 左に行く	-1	0	0
a3: つつく	0	0	**+1**
状態価値	+1/4	+1/2	+1
各状態での最適行動	**a1: 右に行く**	**a1: 右に行く**	**a3: つつく**



結果的にロボットは
右に向いボタンをつつく
行動系列を学習したことになる.

まとめ

1. 強化学習とは**報酬と罰により学習を進める**学習法のことである.
2. 強化学習では**状態と行動**を定義し，それぞれの状態でどの行動をとるべきかを**報酬と価値**に基づいて学習する.
3. **状態行動価値の表**を更新していくことでより多くの報酬を得るための行動が学習できる.
4. 状態の価値を考えずに，すぐに得られる報酬だけを考えて学習したのでは行動を学習することは出来ない.

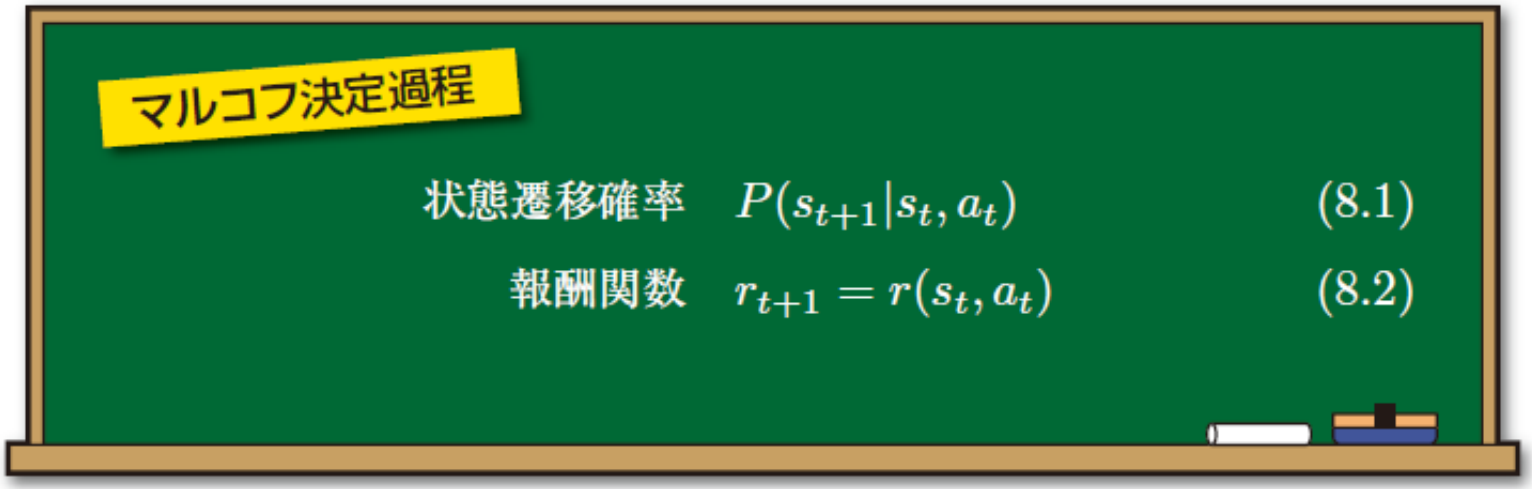
演習8-1 強化学習とは？

- 以下の中で最も**正しくないもの**はどれか？

1. 強化学習とは試行錯誤を通じて報酬を得ながら学習する機械学習のアプローチである．ゆえに教師データは与えられない．
2. 一般的な強化学習のモデルには状態と行動が存在し，それぞれの行動でどのような行動を取るかを学習する．
3. 状態が決まれば通常その後に得られる報酬は一意に決まる．
4. 強化学習は一般的に即時報酬のみで学習するのではなく，遅延報酬も考慮に入れて学習するものである．

Contents

8.2.1 状態遷移確率と報酬関数



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

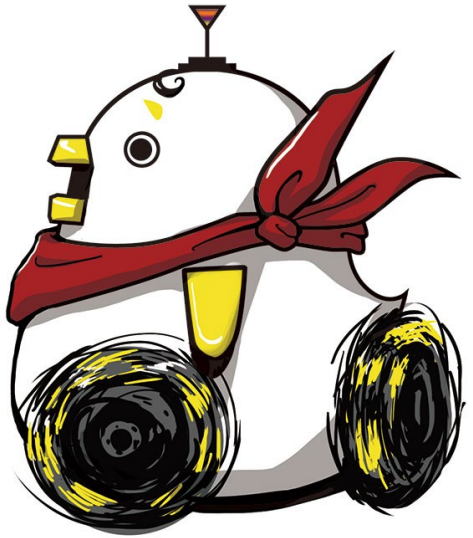
8.2.2 方策と価値

- 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.
  - 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

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

- 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のように「経路」を求めることが問題ではなく，**方策（and/or 価値関数）**を求めることが目的となる.



8.2.3 割引累積報酬の意味

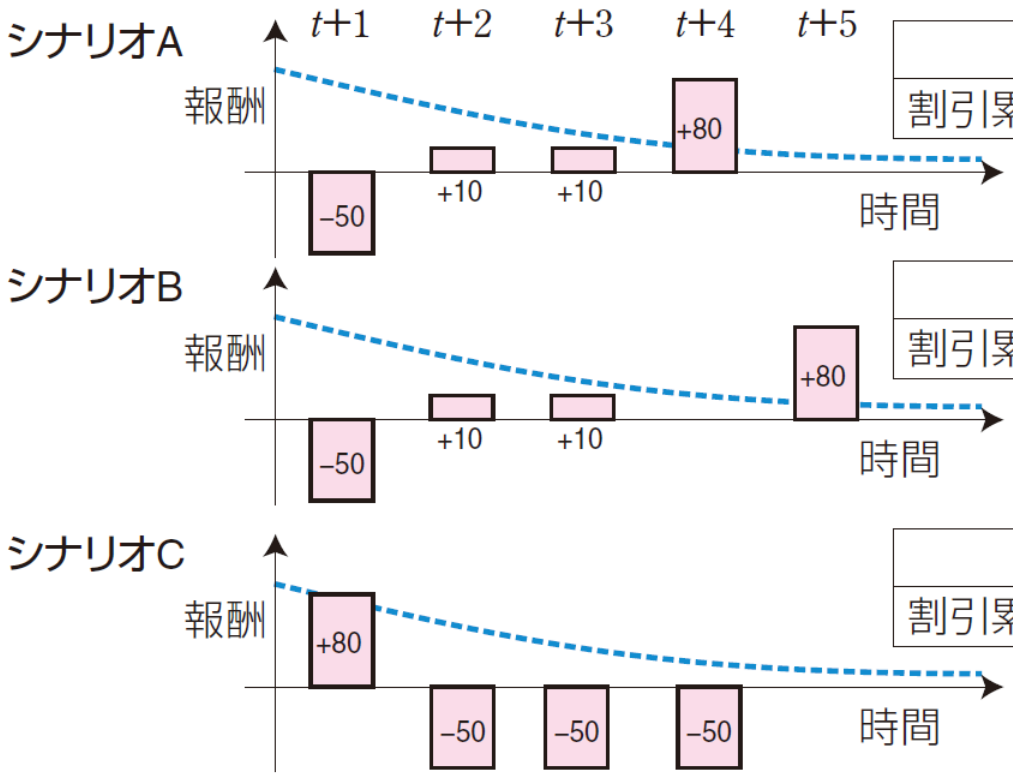
- **割引累積報酬**(discounted return) R_t

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

- 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.
- γ=1 では T→∞で発散する.
- r ：報酬

割引累積報酬の期待値を最大化するような方策を求める

割引率と未来の報酬価値



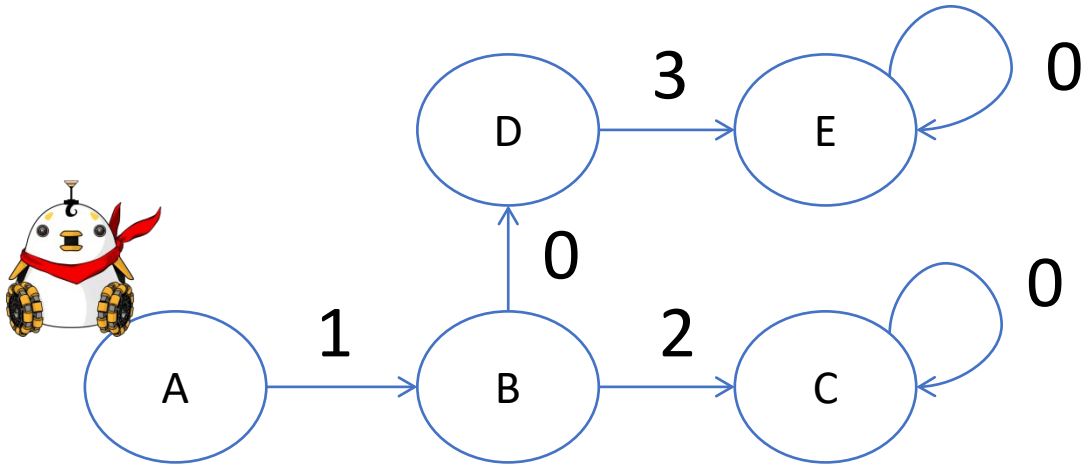
割引率：γ	1.0	0.9	0.5
割引累積報酬：R_t	50	25	−33

割引率：γ	1.0	0.9	0.5
割引累積報酬：R_t	50	20	−38

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

図 8.7　割引累積報酬の割引率による違い

演習8-2割引累積報酬の計算

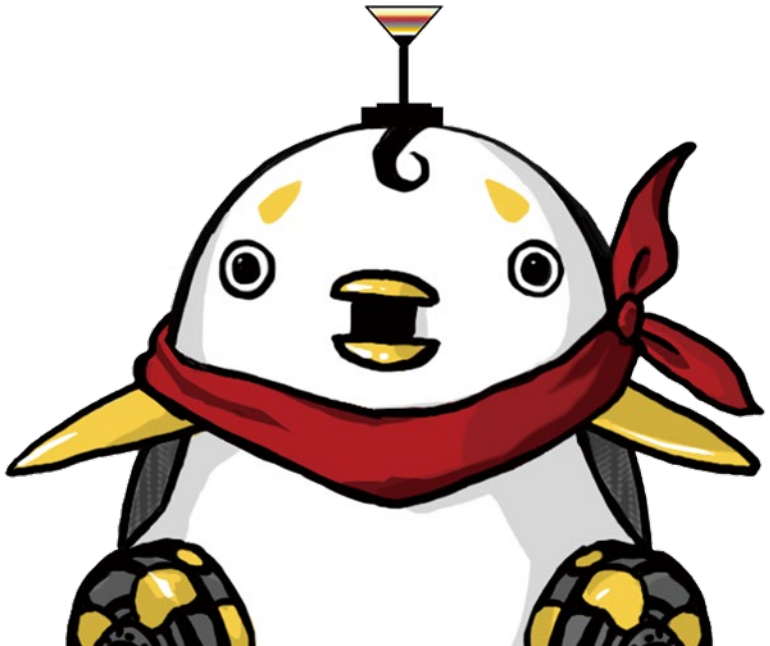


- 方策１は「右へ行けたら右，だめなら上」，方策２は「上へ行けたら上，だめなら右」という方策だとする．両方行けない場合はその場にとどまる．
- 割引率γ= 0.5 の時のA,B,C,D,Eの状態における方策１に従う場合，方策２に従う場合，それぞれで割引累積報酬の値を求めよ．

	A	B	C	D	E
方策１					
方策２					

8.2.4 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

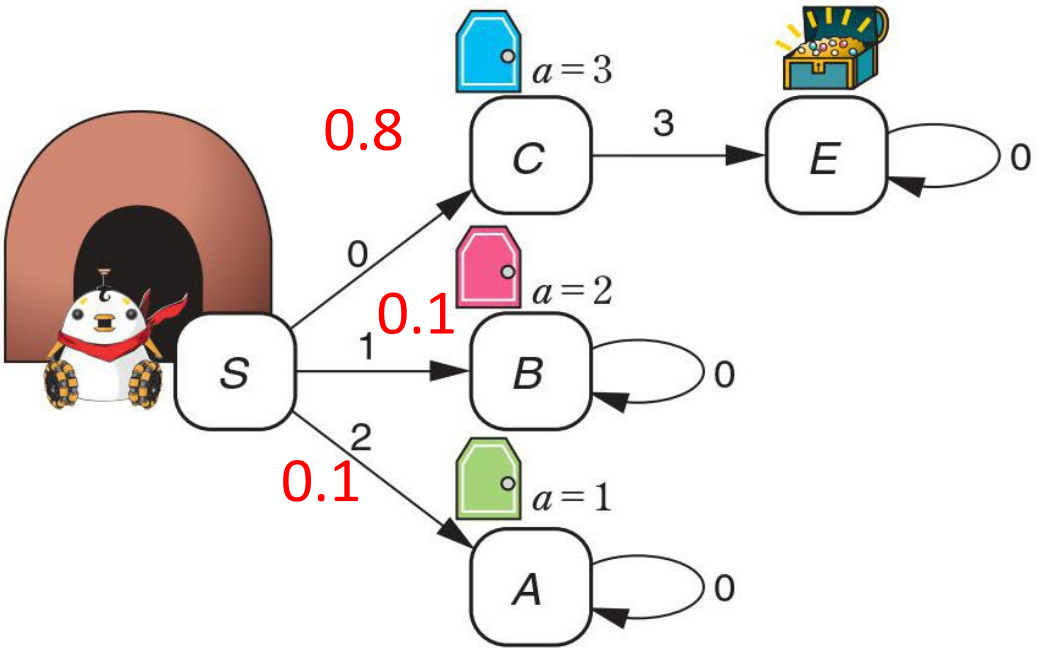
- 8.1 強化学習とは何か？
- 8.2 強化学習の理論
- 8.3 価値関数
- 8.4 学習方法の例：Q 学習

8.3.1 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

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

8.3.2 ホイールダック2 号と分かれ道（確率編）



- 状態遷移確率$P(s_{t+1}|s)$を以下のように定義する。
  S→A：0.1
  S→B：0.1
  S→C：0.8
- 割引率
  γ=0.9

図 8.4 ホイールダック 2 号と分かれ道の先にある報酬

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

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

価値関数の値を高める方策π がよい方策といえる

未来はドンドン分岐する

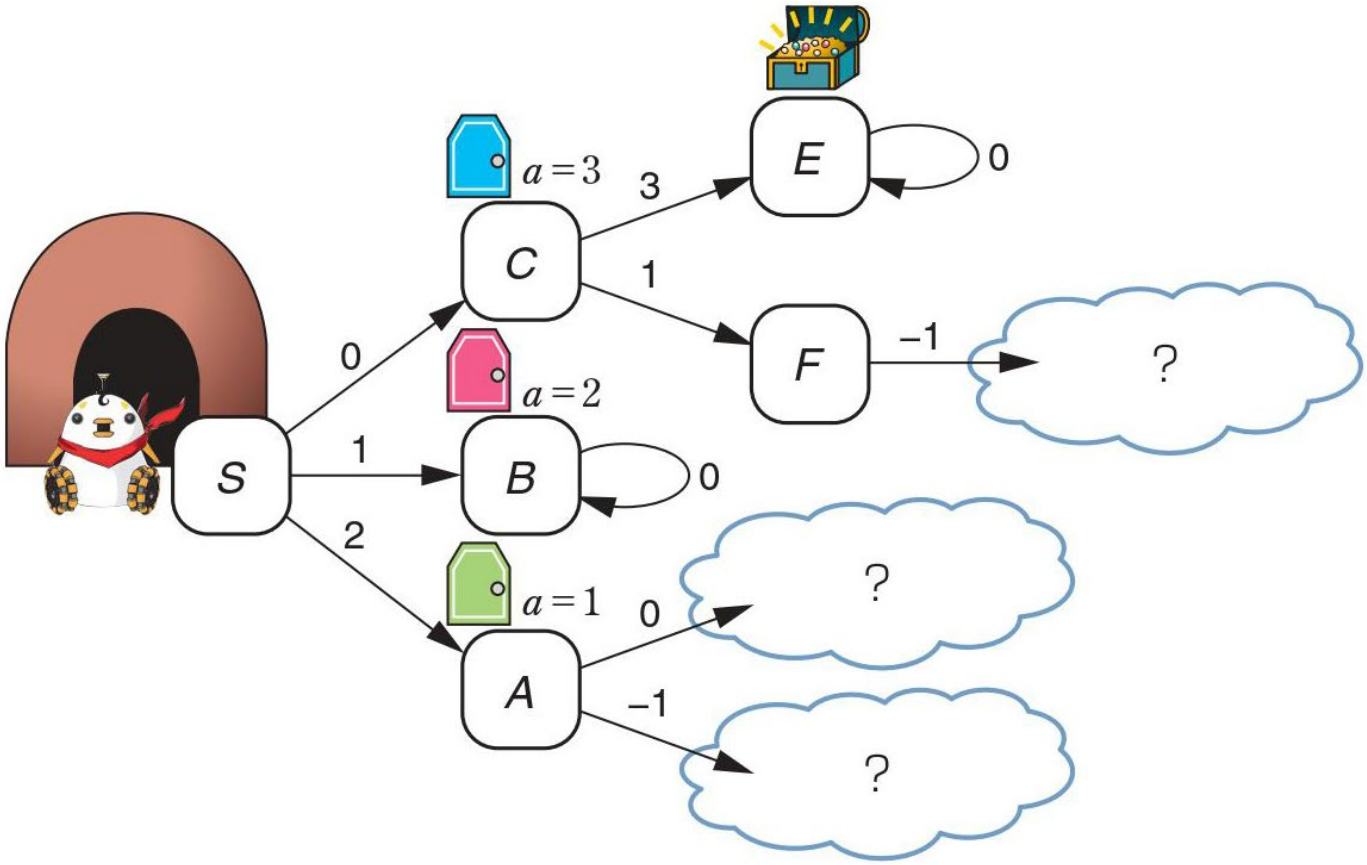
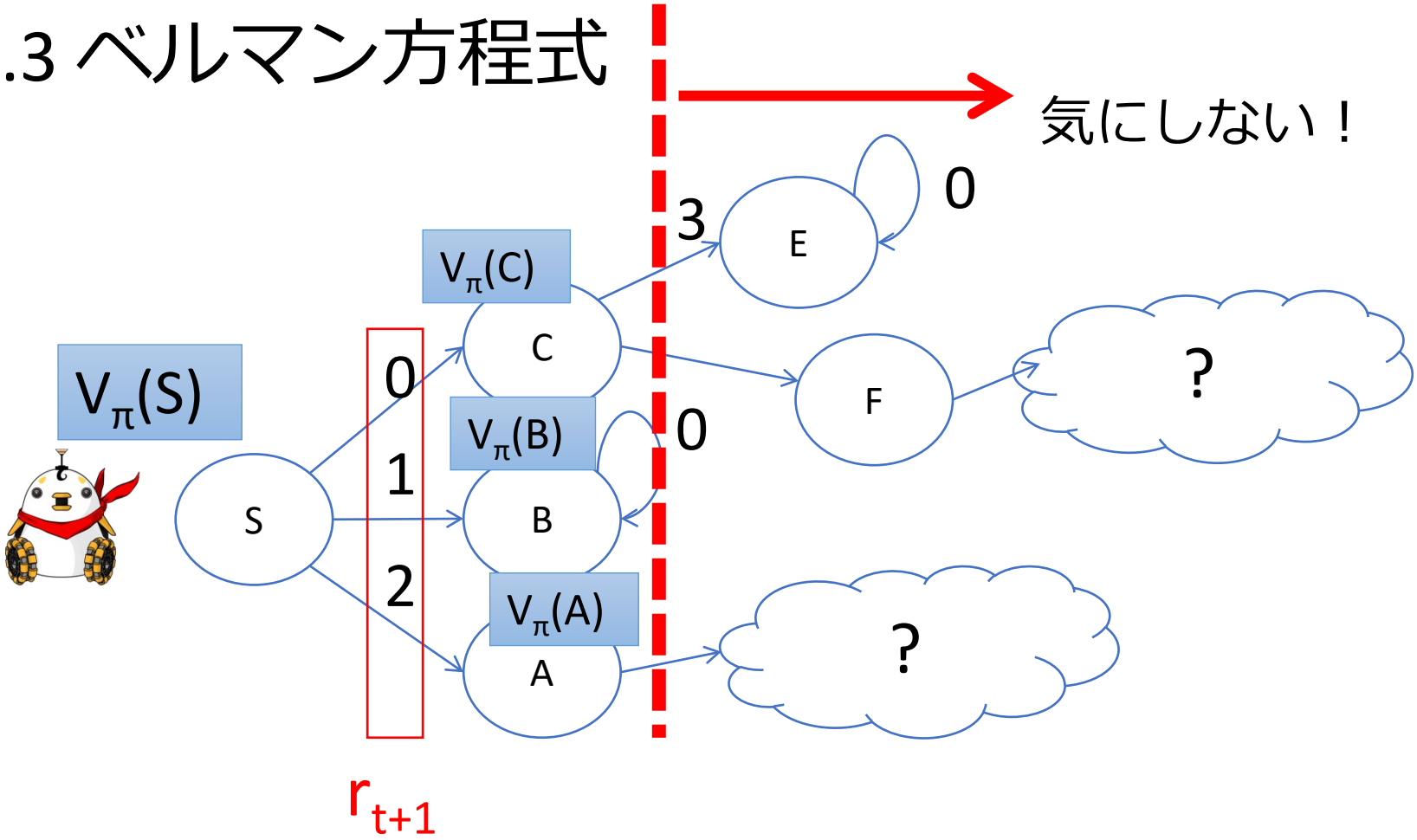


図 8.5　ホイールダック 2 号の未来はどんどん分岐する

□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？

8.3.3 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．

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

8.3.3 行動価値関数

- 行動価値関数(action-value function)

$$Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} V_\pi(s') P(s'|s,a) \tag{8.12}$$

$$V_\pi(s') = \sum_{a'} \pi(s',a') Q_\pi(s',a') \tag{8.13}$$

- 最適行動価値関数

$$Q^*(s,a) := Q_{\pi^*}(s,a) = \max_\pi Q_\pi(s,a) \tag{8.14}$$

行動価値関数の**ベルマン方程式**

$$Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} V_\pi(s') P(s'|s,a) \tag{8.12}$$

$$V_\pi(s') = \sum_{a'} \pi(s',a') Q_\pi(s',a') \tag{8.13}$$

- 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

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

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

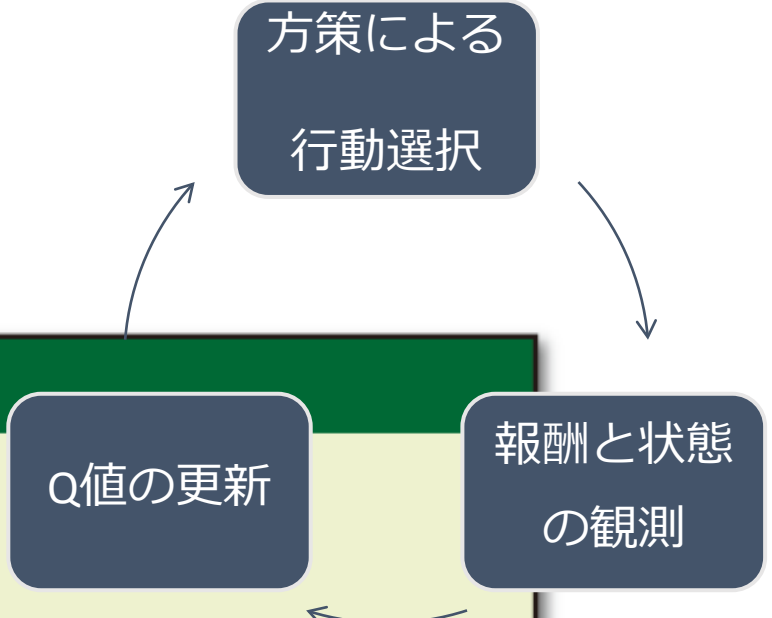
- TD誤差(Temporal difference error)

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

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**

8.4.2 行動選択の方策

- ランダム法
  - 全ての行動を等確率で選択する.
- グリーディ法
  - 各状態においてその時に最適と思われる行動を選択する.
- ε-グリーディ法
  - 確率εでランダムに行動を選択肢，確率(1-ε)でグリーディ法を行う.

$$\pi(s_t, a_t) = P(a_t|s_t) = (1-\varepsilon)\delta(a_t, \underset{a}{\operatorname{argmax}}\, Q(s_t, a)) + \frac{\varepsilon}{\#(A)}$$

- ボルツマン選択
  - パラメータTにより exp(Q(s,a)/T)に比例した確率で行動選択を行う．Tが大きくなればランダム法へ，Tが小さくなればグリーディ法に近づく.

$$\pi(s_t, a_t) = P(a_t|s_t) \propto \exp(\beta Q(s_t, a_t))$$

探索のために「最善でない手」も取らねばならない.

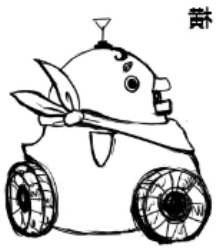
exploration or exploitation trade-off

「知識探索」か「知識活用」か？

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 まとめ

- 試行錯誤から学習する強化学習とは何か理解した.
- 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.
- ϵ-グリーディ法やボルツマン選択といった方策の表現を学んだ.

次回の講義

- 補足説明
- 復習問題
- 解説