

課題3.1

1 レンバイ

S: 部屋のどこにいるかを表す座標，各座標のゴミの量

A: 進む方向

R: 回収したゴミの量

2 赤ちゃんの歩行

S: 間接の角度・速度

A: 間接の角度と速度の変化量

R: 目標キーポイントに進むことができる否か & 安定して姿勢を保つか

(最後の行動報酬も与えても上手く学習できません)

3 魔女

S: 自分の位置・魔女の位置，A: 進む方向，R: 逃げるまでの時間

* 複数step前まで考慮すると、魔女の進行方向が分かり、より上手く逃げれる。

練3.2

部分マニコフ決定過程 (POMDP) で考えべき問題もある。

e.g) 言語生成モデル

環境の状態 + インピュート = s_t
入力文字列

状態 (品詞), 觀測 (單語)

練3.3

環境とインピュートの境界を $\delta = 1 = 31\text{cm}$?

- ・ 入力をモデル化しや可不可以。

練3.4

s	a	s'	r	$p(s', r s, a)$
左	S	左	rs	α
左	S	右	rs	$1 - \alpha$
右	S	左	-3	$1 - \beta$
右	S	右	rs	β
左	W	左	rw	1
左	W	右	rw	1
右	R	左	0	1 //

練3.5

$$\sum_{s' \in S} \sum_{r \in R} p(s', r | s, a) = 1 \quad (\forall s \in S, \forall a \in A(s)) \quad \dots (3.3)$$

↓ イヒヨウ-ドタスク用に変形

S^+ : 非終端状態の集合. S^t : 全ての状態の集合とする.

$$\sum_{s \in S^+} \sum_{r \in R} p(s', r | s, a) = 1 \quad (\forall s \in S, \forall a \in A(s))$$

練3.6

• イヒヨウ-ドタスク

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{T-t-1} R_T$$

$$= -\gamma^{T-t-1} \quad (t < T)$$

• 連続タスク

$$\begin{aligned} G_t &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \\ &= -\gamma^{T_1-t-1} - \gamma^{T_2-t-1} \dots \quad // \\ &\quad (T_1, T_2, \dots \text{は失敗ルート時刻}) \end{aligned}$$

練3.7

エセーリード中の全ての step で収益が 1 に等るので、どの行動が良い行動であるかを学習できません。割引率を使うなどすれば。

練3.8

$$\begin{array}{c} R_1 \qquad \qquad \qquad R_5 \\ -1, 2, 6, 3, 2 \end{array}$$

$$G_5 = 0$$

$$G_4 = R_5 = 2$$

$$G_3 = R_4 + \gamma G_4 = 4$$

$$G_2 = R_3 + \gamma G_3 = 8$$

$$G_1 = R_2 + \gamma G_2 = 6$$

$$G_0 = R_1 + \gamma G_1 = 2$$

練 3.9

$$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \dots$$

$$= \sum_{k=0}^{\infty} \gamma^k r$$

$$= r \cdot \frac{1}{1-\gamma}$$

$$= 70 //$$

$$G_0 = R_1 + \gamma G_1$$

$$= 2 + 63$$

$$= 65 //$$

練 3.10

$$\sum_{k=0}^{\infty} \gamma^k = \lim_{n \rightarrow \infty} \sum_{k=0}^n \gamma^k$$

$$= \lim_{n \rightarrow \infty} \frac{1 - \gamma^{n+1}}{1 - \gamma}$$

$$= \frac{1}{1-\gamma}$$

//

練3.11

$$E_{\pi} [R_{t+1} \mid S_t = s] = \sum_{a \in A(s)} \sum_{\substack{s' \in S \\ r \in R}} r p(s', r \mid s, a) \pi(a \mid s),$$

練3.12

$$\begin{aligned} V^{\pi}(s) &= E_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right] \\ &= \sum_{a \in A(s)} \pi(a \mid s) E_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right] \\ &= \sum_{a \in A(s)} \pi(a \mid s) q^{\pi}(s, a) \end{aligned}$$

練3.13 $p(s', r \mid s, a)$

$$\begin{aligned} q^{\pi}(s, a) &= E_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right] \\ &= \sum_{\substack{s' \in S \\ r \in R}} p(s', r \mid s, a) \left\{ r + E_{\pi} \left[\sum_{k=1}^{\infty} \gamma^k R_{t+k+1} \mid S_{t+1} = s' \right] \right\} \\ &= \sum_{\substack{s' \in S \\ r \in R}} p(s', r \mid s, a) (r + \gamma V^{\pi}(s')) \end{aligned}$$

練 3.14

$$\begin{aligned}
 & \sum_{a \in A} \pi(a|s) \sum_{\substack{s' \in S \\ r \in R}} p(s'|s, a) (r + \gamma V^\pi(s')) \\
 &= \frac{1}{4} (0 + 0.9 \times 2.3 + 0 + 0.9 \times 0.4 + 0 + 0.9 \times (-0.4) + 0 + 0.9 \times 0.7) \\
 &= \frac{2.7}{4} \\
 &= 0.675 \\
 &\doteq 0.7
 \end{aligned}$$

練 3.15

全ての報酬 $r = C$ とする。

$$\begin{aligned}
 V_{\text{new}}^\pi(s) &= E_\pi \left[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + C) \mid S_t = s \right] \\
 &= \frac{C}{1-\gamma} + E_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right] \\
 &= V_C + V^\pi(s)
 \end{aligned}$$

5.2. 相手の価値は変化する。

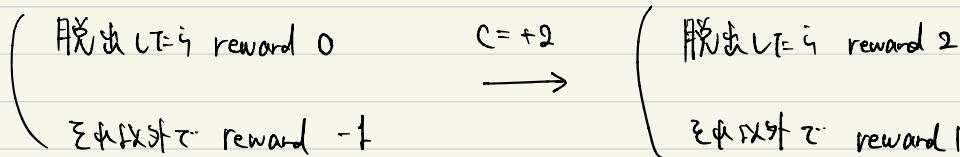
$$V_C = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k C \mid S_t = s \right] = E \left[\sum_{k=0}^{\infty} \gamma^k C \right] = \frac{C}{1-\gamma}$$

4

練3.16

エピソードタスクでは、全ての報酬に C を加えないとタスクは変化する。

e.g) 迷路探索



脱出せずにエピソードを終らせる方向へ incentiveが働くので。

脱出方法を学習しない。エピソードタスクは有限系であるのか

根本の原因

練3.17

$$\begin{aligned} q^{\pi}(s, a) &= E \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \mid S_t = s, A_t = a \right] \\ &= \sum_{\substack{s' \in S \\ r \in R}} p(s', r \mid s, a) \left\{ r + \sum_{a' \in A(s')} \pi(a'|s') E \left[\sum_{k=1}^{\infty} \gamma^k R_{t+k+1} \mid S_{t+1} = s', A_{t+1} = a' \right] \right\} \\ &= \sum_{\substack{s' \in S \\ r \in R}} p(s', r \mid s, a) \left(r + \gamma \sum_{a' \in A(s')} \pi(a'|s') q^{\pi}(s', a') \right) \end{aligned}$$

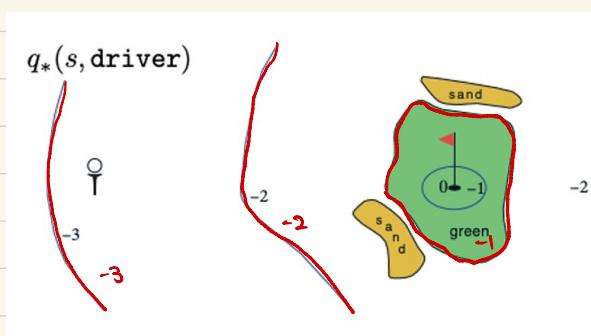
練3.18

$$\begin{aligned} V^\pi(s) &= \sum_{a \in A(s)} \pi(a|s) q^\pi(s, a) \\ &= E_{a \sim \pi(a|s)} [q^\pi(s, a)] // \end{aligned}$$

練3.19

$$\begin{aligned} q^\pi(s, a) &= \sum_{\substack{s' \in S \\ r \in R}} p(s', r | s, a) (r + \gamma V^\pi(s')) \\ &= E_{s', r \sim p(s', r | s, a)} [r + \gamma V^\pi(s')] // \end{aligned}$$

練3.20

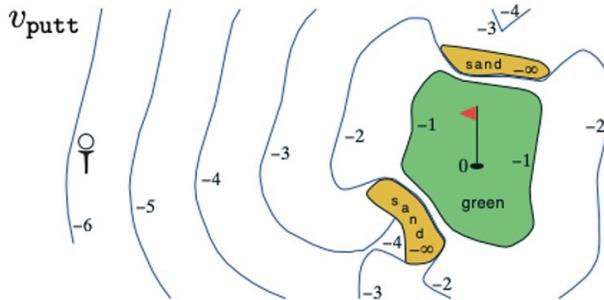


$$V^* = \max_a q^*(s, a)$$

右の -> 内で 1 の -

右の -> 外で 1 の ドライバ -

練3.21

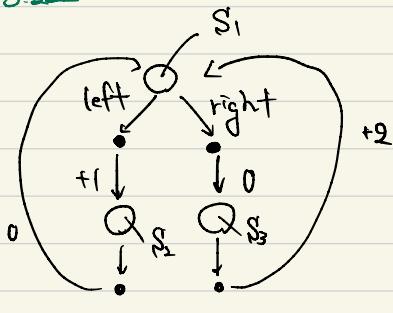


-1, -2, -∞ は上の図と同じ。-3 は, $q^*(s, \text{driver})$ が -2, 等高線内に

putter - FT で属く範囲。-4 は, $q^*(s, \text{driver})$ が -3, 等高線内に putter - FT で属く

範囲なので、下のト位置も含む。

練3.22



π^{left}

$$q^{\pi^{\text{left}}}(s_1, \text{left}) = 1 + \gamma q^V(s_2) = 1 + \gamma (0 + \gamma q^{\pi^{\text{left}}}(s_2, \text{left}))$$

$$\rightarrow q^{\pi^{\text{left}}}(s, \text{left}) = \frac{1}{1 - \gamma^2}$$

$$q^{\pi_{\text{right}}}(s_1, r) = \gamma v(s_3) = \gamma(2 + \gamma q^{\pi_{\text{right}}}(s_3))$$

$$\rightarrow q^{\pi_{\text{right}}}(s, r) = \frac{\gamma}{1 - \gamma^2}$$

5.2. $\gamma = 0.9 \approx 1.0$ π_{left} . $\gamma = 0.5 \approx 1.0$ π_{right} . $\gamma = 0.9 \approx 1.0$ π_{right} 最適

練3.23

$$S = \{h, l\} . A = \{se, wa, re\}$$

$$\text{cf)} \quad q^*(s, a) = \sum_{\substack{s' \in S \\ r \in R}} p(s', r | s, a) (r + \gamma \max_{a'} q^*(s', a'))$$

5.2.

$$q^*(h, se) = \alpha (r_{se} + \gamma \max_{a'} q^*(h, a')) + (1-\alpha) (r_{se} + \gamma \max_{a'} q^*(l, a'))$$

$$q^*(h, wa) = (r_{wa} + \gamma \max_{a'} q^*(h, a'))$$

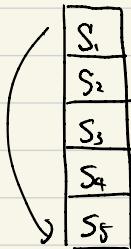
$$q^*(l, se) = \beta (r_{se} + \gamma \max_{a'} q^*(l, a')) + (1-\beta) (-3 + \gamma \max_{a'} q^*(h, a'))$$

$$q^*(l, wa) = (r_{wa} + \gamma \max_{a'} q^*(h, a'))$$

$$q^*(l, re) = \gamma \max_{a'} q^*(h, a') \quad //$$

練 3.24

$$\text{cf. } V^*(s) = \max_a \sum_{s', r} p(s', r | s, a) (r + \gamma V^*(s'))$$



$$\begin{aligned} V^*(s_1) &= 10 + \gamma V^*(s_5) \\ &= 10 + \gamma (10 + \gamma V^*(s_4)) \\ &\quad \vdots \\ &= 10 + \gamma^5 V^*(s_1) \end{aligned}$$

$$V^*(s_1) = \frac{10}{1 - \gamma^5} = \frac{10}{1 - 0.95} = 204.19 //$$

練 3.25

$$V^*(s) = \max_a q^*(s, a) //$$

練 3.26

$$q^*(s, a) = \sum_{\substack{s' \in S \\ r \in R}} p(s', r | s, a) (r + \gamma V^*(s')) //$$

練 3.27

$$\pi^*(a | s) = \max_a q^*(s, a) //$$

練 3.28

$$\pi^*(a | s) = \max_a \sum_{\substack{s' \in S \\ r \in R}} p(s', r | s, a) (r + \gamma V^*(s')) //$$

練3.29 $p(s'|s,a) + p(r|s,a)$ を用いてベルマン方程式を記述 //

この式が「 $p(s',r|s,a)$ 」の定義式である。 //

$$\begin{aligned} \circ v^\pi(s) &= \sum_{a \in A(s)} \pi(a|s) \sum_{\substack{s' \in S \\ r \in R}} p(s',r|s,a) (r + \gamma v^\pi(s')) \\ &= \sum_{a \in A(s)} \pi(a|s) \left\{ \sum_{r \in R} r p(r|s,a) + \sum_{s' \in S} \gamma v^\pi(s') p(s'|s,a) \right\} \end{aligned}$$

$$\begin{aligned} \circ q^\pi(s,a) &= \sum_{\substack{s' \in S \\ r \in R}} p(s',r|s,a) (r + \gamma \sum_{a' \in A(s')} \pi(a'|s') q^\pi(s',a')) \\ &= \sum_{r \in R} r p(r|s,a) + \sum_{s' \in S} p(s'|s,a) \left\{ \gamma \sum_{a' \in A(s')} \pi(a'|s') q^\pi(s',a') \right\} \end{aligned}$$

$$\begin{aligned} \circ v^*(s) &= \max_a \sum_{\substack{s' \in S \\ r \in R}} p(s',r|s,a) (r + \gamma v^*(s')) \\ &= \max_a \left\{ \sum_{r \in R} r p(r|s,a) + \sum_{s' \in S} \gamma v^*(s') p(s'|s,a) \right\} \end{aligned}$$

$$\begin{aligned} \circ q^*(s,a) &= \sum_{\substack{s' \in S \\ r \in R}} p(s',r|s,a) (r + \gamma \max_{a'} q^*(s',a')) \\ &= \sum_{r \in R} r p(r|s,a) + \sum_{s' \in S} p(s'|s,a) \gamma \max_{a'} q^*(s',a') \end{aligned}$$

//