

練4.1

$$q_{\pi}(s,a) = \sum_{\substack{s' \in S \\ r \in R}} p(s',r|s,a) (r + \gamma V_{\pi}(s'))$$

$$g_{\pi}(11, \text{down}) = 1(-1 + 7 \cdot 0) = -1$$

$$g_{\pi}(7, \text{down}) = 1(-1 + g(-14)) = -15$$

練 4.2

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	
15			

$$V_{\pi}(15) = \sum_{a \in A(15)} \sum_{\substack{s' \in S \\ r \in R}} p(s', r | s, a) (r + \gamma V_{\pi}(s'))$$

$$= \frac{1}{4} (-4 + V_{\pi}(15) - 22 - 20 - 14)$$

$$U_{\pi}(15) = \frac{4}{3} \cdot \frac{1}{4}(-60) = -20 //$$

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	
15			

$$U_{\pi}(13) = -1 + \frac{1}{2} (U_{\pi}(9) + U_{\pi}(15) + 9U_{\pi}(12) + 9U_{\pi}(14))$$

$$V_{\pi}(15) = -1 + \frac{1}{4} (V_{\pi}(13) + V_{\pi}(15) + V_{\pi}(12) + V_{\pi}(14))$$

图 4.1 の $k = \infty$ の V は、 $V(15) = -20$ エビデンスも見て初期値をみて

1 step の更新で v_t は変化しない。 $\delta_t = V_{\pi}(s_t) - v_t$

練4.3

$$q_{\pi}(s, a) = E \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right]$$

$$= \sum_{\substack{s' \in S \\ r \in R}} p(s', r \mid s, a) (r + \gamma \sum_{a' \in A(s')} \pi(a' \mid s') q_{\pi}(s', a'))$$

$$q_{K+1}(s, a) = \sum_{\substack{s' \in S \\ r \in R}} p(s', r \mid s, a) (r + \gamma \sum_{a' \in A(s')} \pi(a' \mid s') q_K(s', a')) \quad //$$

練4.4

最適方策の複数ある時、最適状態価値関数は一意。よって方策改善

はおこる。 $\pi_{K+1}(s) \in q_{\pi_K}(s, a)$ を最大にできる a のうち最小の r とすればよい。

↓
a1. 順序で最大化。

練4.5

$$\pi_0 \rightarrow q_{\pi_0} \rightarrow \pi_1 \rightarrow q_{\pi_1} \rightarrow \dots$$

(i) 方策評価 $\pi_K \rightarrow q_{\pi_K}$

$$q_{\pi_K}(s, a) \leftarrow \sum_{\substack{s' \in S \\ r \in R}} p(s', r \mid s, a) (r + \gamma \sum_{a' \in A(s')} \pi(a' \mid s') q_{\pi_K}(s', a'))$$

を収束まで行う。

(ii) 方界改善 $q_{\pi_k} \rightarrow \pi_{k+1}$

$$\pi_{k+1}(s) = \underset{a}{\operatorname{argmax}} q_{\pi_k}(s, a)$$

練 4.6

3. 方界改善 ... $\pi(a|s) = \begin{cases} 1 - \varepsilon & (a = \underset{a'}{\operatorname{argmax}} q_{\pi}(s, a')) \\ \frac{\varepsilon}{|A|} & (\text{otherwise}) \end{cases}$

2. 右界評価 ... $V(s) \leftarrow \sum_{a \in A(s)} \pi(a|s) \sum_{\substack{s' \in S \\ r \in R}} p(s'|r|sa) (r + \gamma V(s'))$

1. 初期化 ... $\pi(a|s)$ は任意、確率分布で初期化可

練 4.7

$$S = \{(n_1, n_2) \mid n_1, n_2 \in \mathbb{N}, n_1 + n_2 \leq 20\}$$

$$A(n_1, n_2) = \{x \mid -\min(5, n_2) \leq x \leq \min(5, n_1)\}$$

↑
1 → 2 と車の数

$$S_t \xrightarrow{k} \underbrace{\quad}_{\text{車移動}} \xrightarrow{k} S_{t+1}$$

車移動
車の増減

```

Iter 1:
Policy Evaluation
# of iteration: 72
Policy Improvement
371 / 441 are updated
Iter 2:
Policy Evaluation
# of iteration: 57
Policy Improvement
265 / 441 are updated
Iter 3:
Policy Evaluation
# of iteration: 49
Policy Improvement
102 / 441 are updated
Iter 4:
Policy Evaluation
# of iteration: 32
Policy Improvement
0 / 441 are updated
[ 0 0 0 0 0 0 0 -1 -1 -2 -2 -3 -3 -3 -4 -5 -4 -4 -4 -5 -5 -5 -5 ]
[ 1 0 0 0 0 0 0 0 -1 -1 -2 -2 -3 -3 -3 -4 -5 -3 -4 -4 -4 -4 -4 ]
[ 1 1 0 0 0 0 0 0 0 -1 -1 -1 -2 -3 -4 -5 -3 -3 -3 -3 -3 -3 ]
[ 1 1 1 1 0 0 0 0 0 0 0 0 -1 -2 -3 -4 -5 -2 -2 -2 -2 -2 ]
[ 1 1 1 1 1 0 0 0 0 0 0 0 -1 -2 -3 -4 -1 -1 -1 -1 -1 -1 ]
[ 1 1 1 1 1 1 0 0 0 0 0 0 -1 -2 -3 0 0 0 0 0 0 0 -1 ]
[ 2 1 1 1 1 1 1 0 0 0 0 -1 -2 0 0 0 0 0 0 0 0 0 0 ]
[ 2 2 1 1 1 1 1 1 1 0 0 -1 -2 0 0 0 0 0 0 0 0 0 0 ]
[ 3 2 2 1 1 1 1 1 1 1 0 -1 0 0 0 0 0 0 0 0 0 0 0 ]
[ 4 3 3 2 1 1 1 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 ]
[ 4 4 3 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 ]
[ 5 4 3 3 2 2 2 2 1 0 2 2 2 2 2 2 2 2 2 2 1 0 0 ]
[ 5 4 4 3 3 3 3 3 1 1 0 -1 3 3 3 3 3 3 3 1 1 0 ]
[ 5 5 4 4 4 4 4 4 1 1 1 0 -1 1 1 1 1 1 1 1 1 1 0 ]
[ 5 5 5 5 5 5 1 1 1 1 0 -1 1 1 1 1 1 1 1 1 1 0 ]
[ 5 5 4 4 3 2 1 1 1 1 0 -1 1 1 1 1 1 1 1 1 1 0 ]
[ 5 5 5 4 3 2 1 1 1 1 0 -1 1 1 1 1 1 1 1 1 1 0 ]
[ 5 5 5 4 3 2 2 1 1 1 0 -1 1 1 1 1 1 1 1 1 1 0 ]
[ 5 5 5 4 3 3 2 1 1 1 0 -1 1 1 1 1 1 1 1 1 1 0 ]
[ 5 5 5 4 4 3 2 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 ]

```

1344.2

4.7

練4.8

$p_h < 0.5$ の時、step 数をでまかせるのがいいらしい。 $\delta = 50\%$ ときは、全額 bet で 勝率が 4割である。二つ以上の勝率は達成できない。 $\delta = 5\%$ で bet の最適な理由は不明…。

練4.9 手牌の png file 参照

- $\theta \rightarrow 0$ のとき $p_h = 0.25$ のときは 不安定。 $p_h = 0.55$ のときは 安定。

練4.10

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

(↑)

$$\pi(s) = \max_a q(s, a)$$

\downarrow これはまとめて

$$q_{k+1}(s, a) = \sum_{\substack{s' \in S \\ r \in R}} p(s', r | s, a) (r + \gamma \max_{a'} q_k(s', a')) \quad //$$