Q-学習の更新式は以下のとおり。

$$Q(S_t, A_t) \leftarrow 0.9\,Q(S_t, A_t) + 0.1\left(R_{t+1} + 0.9 \max Q(S_{t+1}, a)\right)$$

この式を元に行動価値関数を求めていく。これにより、

(1, R) → (2, R) と行動すればよいということが分かる。

1 $Q(S_2, R) = 0.9 \cdot 0 + 0.1(100 + 0.9 \cdot 0) = 10$

2 $Q(S_1, R) = 0.9 \cdot 0 + 0.1(0 + 0.9 \times 10) = 0.9$

3 $Q(S_1, L) = 0.9 \times 0 + 0.1 \cdot 0 = 0$

4 $Q(S_2, R) = 0.9 \times 10 + 0.1(100 + 0.9 \cdot 0) = 19$

5 $Q(S_1, L) = 0.9 \times 0 + 0.1 \cdot 0 = 0$

6 $Q(S_1, R) = 0.9 \times 0.9 + 0.1(0 + 0.9 \times 9) = 2.52$

7 $Q(S_2, L) = 0.9 \times 0 + 0.1(0 + 0.9 \times 2.52) = 0.227$

8 $Q(S_1, R) = 0.9 \times 2.52 + 0.1(0.9 \times 19) = 3.978$

9 $Q(S_2, R) = 0.9 \times 19 + 0.1(100 + 0) = 27.1$

t	S_t a_t	1L	1R	2L	2R
0	1 R	0	0	0	0
1	2 R	0	0	0	10
2	1 L	0	0	0	10
3	1 R	0	0.9	0	10
4	2 R	0	0.9	0	19
5	1 L	0	0.9	0	19
6	1 R	0	2.52	0	19
7	2 L	0	2.52	0.227	19
8	1 R	0	3.978	0.227	19
9	2 R	0	3.978	0.227	27.1