Machine Learning

Assignment 11.3

Submitted By: Ranji Raj

January 23, 2021

a) Value-Iteration, Q-table, Optimal policy, $\gamma = 0.8$

Algorithm

- Start with $V(s) \leftarrow \underset{a}{max} \quad r(s,a), \forall s$
- Until V changes, perform for all states s,

$$V(s) \leftarrow \underset{a}{max}\{r(s,a) + \gamma V(\delta(s,a))\}$$

- Choose the optimal policy:

$$\pi(s) = \arg\max_a \{r(s,a) + \gamma V(\delta(s,a))\}$$

$$V^*(s) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + ... \equiv \sum_{i=0}^{\infty} \gamma^i r_t + i$$

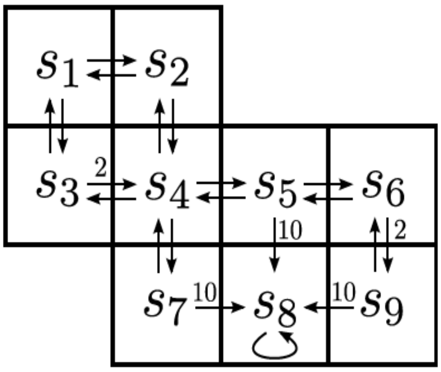
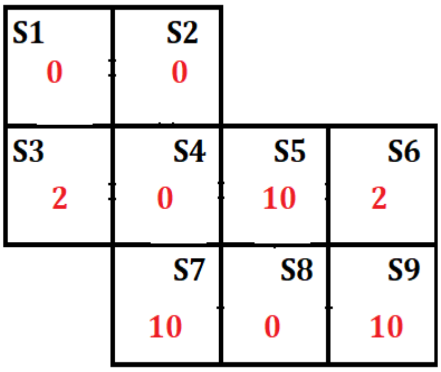


Figure 1: Deterministic grid world

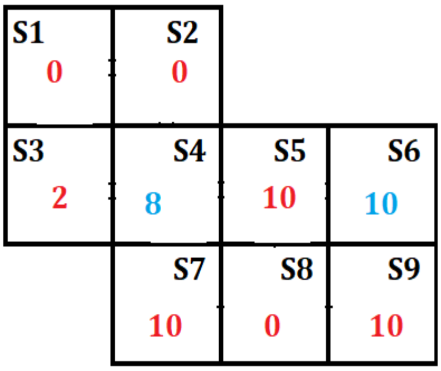


S1 0	S2 0		
S3 2	S4 0	S5 10	S6 2
	S7 10	S8 0	S9 10

Figure 2: Initialization

state	**calculation**	V^*
S_6	2+(0.8)(10)	10
S_4	0+(0.8)(10)	8

Table 1: Iteration-1

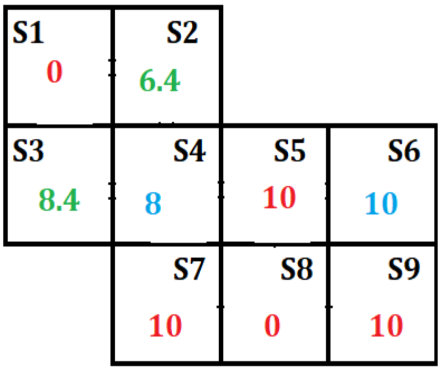


S1 0	**S2** 0		
S3 2	**S4** 8	**S5** 10	**S6** 10
	S7 10	**S8** 0	**S9** 10

Table 2: Grid-1

state	**calculation**	V^*
S_3	$2+(0.8)(0)+(0.8)^2(10)$	8.4
S_2	$0+(0.8)(0)+(0.8)^2(10)$	6.4

Table 3: Iteration-2

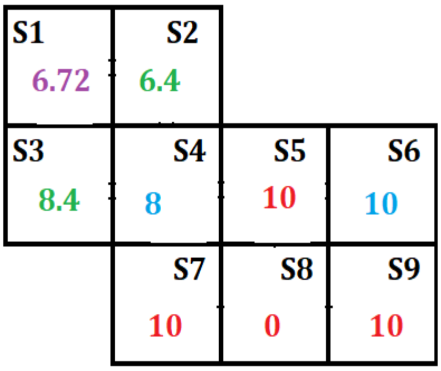


S1 0	**S2** 6.4		
S3 8.4	**S4** 8	**S5** 10	**S6** 10
	S7 10	**S8** 0	**S9** 10

Table 4: Grid-2

state	**calculation**	V^*
S_1	$0+(0.8)(2)+(0.8)^2(0)+$ $(0.8)^3(10)$	6.72

Table 5: Iteration-3



S1 6.72	**S2** 6.4		
S3 8.4	**S4** 8	**S5** 10	**S6** 10
	S7 10	**S8** 0	**S9** 10

Table 6: Grid-3

S_{start}	S_{stop}	**calculation** $r + \gamma V^*(S_{stop})$	$Q(s,a)$
S_1	S_2	$0 + 0.8(6.4)$	5.12
S_1	S_3	$0 + 0.8(8.4)$	**6.72**
S_2	S_1	$0 + 0.8(6.72)$	5.37
S_2	S_4	$0 + 0.8(8)$	**6.4**
S_3	S_1	$0 + 0.8(6.72)$	5.37
S_3	S_4	$2 + 0.8(8)$	**8.4**
S_4	S_2	$0 + 0.8(6.4)$	5.12
S_4	S_3	$0 + 0.8(8.4)$	6.72
S_4	S_5	$0 + 0.8(10)$	**8.00**
S_4	S_7	$0 + 0.8(10)$	**8.00**
S_5	S_5	$0 + 0.8(8)$	6.4
S_5	S_6	$0 + 0.8(10)$	8.00
S_5	S_8	$10 + 0.8(0)$	**10.00**
S_6	S_5	$0 + 0.8(10)$	8.00
S_6	S_9	$2 + 0.8(10)$	**10.00**
S_7	S_4	$0 + 0.8(8)$	6.4
S_7	S_8	$10 + 0.8(0)$	**10.00**
S_9	S_6	$0 + 0.8(10)$	8.00
S_9	S_8	$10 + 0.8(0)$	**10.00**

Table 7: Q-table

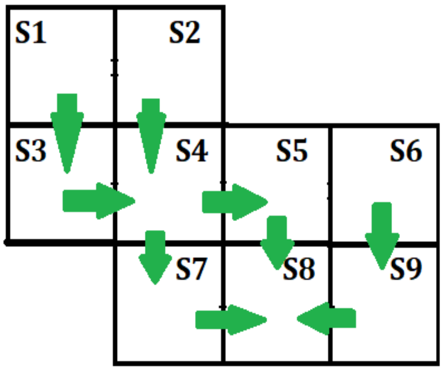


Table 8: Optimal policy

b) Modifying reward function r(s,a)

Alters Q(s,a) but not optimal policy

Multiply every reward with a constant value (say, 10). Rewards will change which in turn changes Q(s,a) but optimal policy won't be affected.

Alters Q(s,a) but not V^*

Choose a reward $r(S_5, S_6) = 1$
Direction $\rightarrow$: $V^* = 1 + (0.8)2 + (0.8)^2 10 = 9$
Direction $\downarrow$: $V^* = 10 + (0.8)0 = 10 \Rightarrow$ still the best V^*