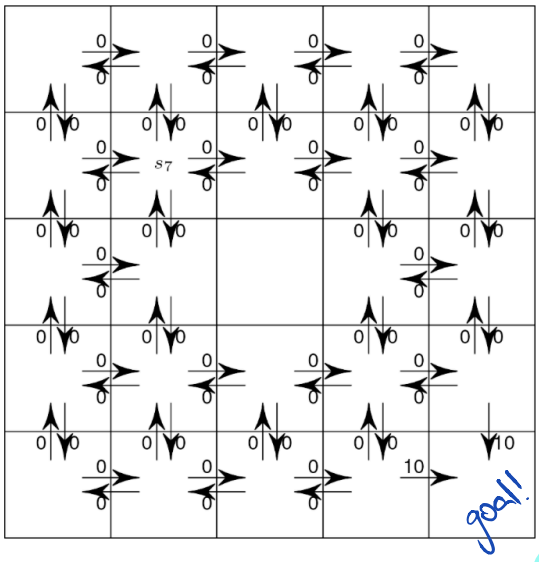
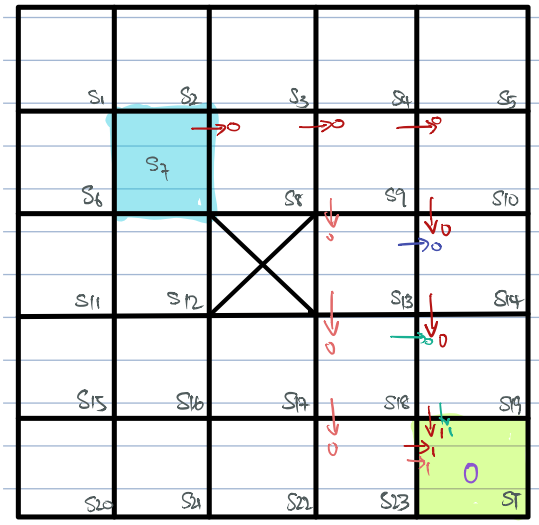
1. Consider the gridworld shown in Figure 1 with a goal state in the lower right-hand corner. Show (using a drawing or a table) an optimal policy for environment and calculate the value $V^*(s_7)$ of the state s_7 assuming the discounting factor $\gamma = 0.8$. (*Answer*: 3.2768)





All actions have 0 rewards, except for the immediate action that leads to S_T, i.e.

(s,a)	r
$s_{19}, \downarrow$	10
$s_{19}, \uparrow$	0
$s_{19}, \leftarrow$	0

(s,a)	r
$s_{23}, \rightarrow$	10
$s_{23}, \uparrow$	0
$s_{23}, \leftarrow$	0

i.e. optimal policy is the one with the shortest route! & there are 4 such paths that exist!

Bellman's eqn: $V_s = \max_{a \in A} \pi(a|s) \cdot (r_{s,a} + \gamma V_{s'})$

Figure 1: A gridworld with a goal state.

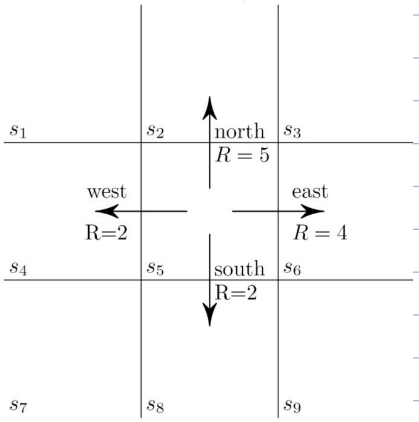


Figure 2: A fragment of a gridworld.

$V_7^* = \gamma V_8^*$ $V_8^* = \gamma V_9^*$ $V_9^* = 0.5\gamma[V_{10}^* + V_{13}^*]$ $V_{10}^* = \gamma V_{14}^*$

$V_{13}^* = 0.5\gamma[V_{18}^* + V_{14}^*]$

$V_{14}^* = \gamma V_{19}^*$ $V_{18}^* = 0.5\gamma[V_{19}^* + V_{23}^*]$ $V_{19}^* = 10$ $V_{23}^* = 10$

$\therefore V_{18}^* = 0.5(0.8)[10+10] = 8$

$V_{14}^* = 0.8$

$V_{10}^* = 0.8[0.8] = 6.4$

$V_{13}^* = 0.5(0.8)[8+8] = 6.4$

$V_9^* = 0.5(0.8)[6.4+6.4] = 5.12$

$V_8^* = 0.8(5.12) = 4.096$

$V_7^* = 0.8(4.096) = 3.2768$ #

2. Explain, in your own words, what is the difference between a deterministic and a stochastic policy.

3. Consider the fragment of a gridworld in Figure 2. Assuming the discounting factor $\gamma = 0.8$, write down the Bellman equations for the policy evaluation at the state s_5 for the following policies:

   (a) deterministic policy with $\pi_1(s_5) = \text{east}$;

   (b) stochastic policy with the following distribution on actions at the state s_5:

action a	$\pi_2(a \mid s_5)$
north	0.4
east	0.2
south	0.1
west	0.3

2) Deterministic policy refers to policy where there is only 1 possible action at a state

i.e. $\pi(a_n|s) = 1$ & $\pi(a_m|s) = 0$ for $a \in \{a_a \dots a_n, a_m, a \dots\}$ where $n \neq m$.

Stochastic policy however, allows the agent choose between more than 1 action at a given state i.e. $0 \leq \pi(a_i|s) \leq 1$

s.t. $\sum_{n=1}^{N} \pi(a_i|s) = 1$.

3. $V_s = \max_{a \in A} \pi(a|s)[r_{s,a} + \gamma V_{s'}]$

a) $\pi_1(S_5) = \text{east}$ & $r_{east} = 4 \Rightarrow V_\pi(S_5) = 1 \cdot [4 + \gamma V_\pi(S_6)] = 4 + 0.8 V_\pi(S_6)$

b) $V_\pi(S_5) = 0.4 \cdot [5 + \gamma V_\pi(S_2)] + 0.2 \cdot [4 + \gamma V_\pi(S_6)] + 0.1 \cdot [2 + \gamma V_\pi(S_8)] + 0.3 \cdot [2 + \gamma V_\pi(S_4)]$

$= 2 + 0.8 + 0.2 + 0.6 + 0.32 V_\pi(S_2) + 0.16 V_\pi(S_6) + 0.08 V_\pi(S_8) + 0.24 V_\pi(S_4)$

$= 3.6 + 0.32 V_\pi(S_2) + 0.16 V_\pi(S_6) + 0.08 V_\pi(S_8) + 0.24 V_\pi(S_4)$

4. Consider the fragment of a gridworld in Figure 2. Let Table 1 give estimates of a policy values. Assuming the discounting rate of $\gamma = 0.8$, use the Bellman equation to update the estimate of the value of s_5

   (a) if the policy is π_1 from above (*Answer:* 10.4);

   (b) if the policy is π_2 from above (*Answer:* 7.68).

5. Consider the fragment of a gridworld in Figure 2. Let Table 1 give policy values $V_\pi(s)$. Assuming the discounting rate of $\gamma = 0.8$, calculate the values of Q-function for s_5.

Table 1: Estimates of state values

state	estimate
s_1	2
s_2	4
s_3	8
s_4	5
s_5	4
s_6	8
s_7	4
s_8	4
s_9	4

$V_S = \sum_{a \in A} \pi(a|S) \cdot [r_{s,a} + \gamma V_{S'}]$

a) $V_{\pi_1}(S_5) = 4 + 0.8 V_{\pi_1}(S_6) = 4 + 0.8(8) = 4 + 64 = 104$ ✗

b) $V_{\pi_2}(S_5) = 3.6 + 0.32 V_\pi(S_2) + 0.16 V_\pi(S_6) + 0.08 V_\pi(S_8) + 0.24 V_\pi(S_4)$
$= 3.6 + 0.32(4) + 0.16(8) + 0.08(4) + 0.24(5)$
$= 7.68$ ✓

6. Consider the fragment of a gridworld in Figure 2. Let Table 1 give estimates of optimal values $V^*(s)$. Update the estimate of $V^*(s_5)$ using the Bellman equation. (*Answer:* 10.4.)

$V^*(S_5) = \max_{a \in A} [r_{s,a} + \gamma V^*(S')]$

$a = \uparrow: r_{5,\uparrow} + \gamma V^*(S_2) = 5 + 0.8(4) = 8.2$
$a = \leftarrow: r_{5,\leftarrow} + \gamma V^*(S_4) = 2 + 0.8(5) = 6$
$a = \rightarrow: r_{5,\rightarrow} + \gamma V^*(S_6) = 4 + 0.8(8) = 10.4$
$a = \downarrow: r_{5,\downarrow} + \gamma V^*(S_8) = 2 + 0.8(4) = 5.2$

$\therefore V^*(S_5) = 10.4$

7. Consider a fragment of a stochastic MDP in Figure 3. In this process, an action may lead to different rewards and states with some probabilities; this is shown by splitting arrows with probabilities and rewards written on them. Assuming the discounting factor $\gamma = 0.8$, write down the Bellman equations for the policy evaluation at the state s_1 for the following policies:

   (a) deterministic policy with $\pi_3(s_1) = a_2$;

   (b) stochastic policy with

action a	$\pi_4(a \mid s_1)$
a_1	0.4
a_2	0.2
a_3	0.4

(*Answers:* $V_{\pi_3}(s_1) = 1.5 + 0.4V_{\pi_3}(s_3) + 0.4V_{\pi_3}(s_4)$; $V_{\pi_4}(s_1) = 2.38 + 0.32V_{\pi_4}(s_2) + 0.08V_{\pi_4}(s_3) + 0.208V_{\pi_4}(s_4) + 0.192V_{\pi_4}(s_5)$.)

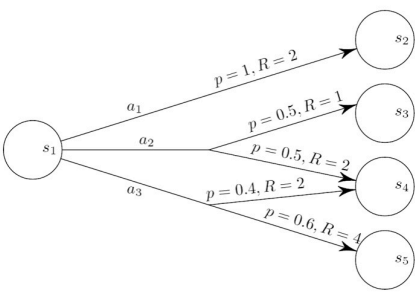


Figure 3: A fragment of a stochastic MDP.

$V_\pi(S) = \sum_{a \in A} \pi(a|S) [r_{s,a} + \gamma V(S')]$

a) $\pi_3(S_1) = a_2$

$\therefore V_{\pi_3}(S_1) = 0.5[1 + 0.8 V_{\pi_3}(S_3)] + 0.5[2 + 0.8 V_{\pi_3}(S_4)]$
$= 0.5 + 0.4 V_{\pi_3}(S_3) + 1 + 0.4 V_{\pi_3}(S_4)$
$= 1.5 + 0.4 V_{\pi_3}(S_3) + 0.4 V_{\pi_3}(S_4)$.

b) $V_{\pi_4}(S_1) = 0.4 \times [2 + 0.8 V_{\pi_4}(S_2)] + 0.2 \times [0.5 \times [1 + 0.8 V_{\pi_4}(S_3)] + 0.5 \times [2 + 0.8 V_{\pi_4}(S_4)]] + 0.4 \times [0.4 \times [2 + 0.8 V_{\pi_4}(S_4)] + 0.6 \times [4 + 0.8 V_{\pi_4}(S_5)]]$
$= 0.8 + 0.32 V_{\pi_4}(S_2) + 0.2[0.5 + 0.4 V_{\pi_4}(S_3) + 1 + 0.4 V_{\pi_4}(S_4)] + 0.4 \times [0.8 + 0.32 V_{\pi_4}(S_4) + 2.4 + 0.48 V_{\pi_4}(S_5)]$
$= 2.38 + 0.32 V_{\pi_4}(S_2) + 0.08 V_{\pi_4}(S_3) + 0.208 V_{\pi_4}(S_4) + 0.192 V_{\pi_4}(S_5)$ ✓

8. Consider the fragment of a gridworld in Figure 3. Let Table 1 give estimates of a policy values. Assuming the discounting rate of $\gamma = 0.8$, use the Bellman equation to update the estimate of the value of s_1

   (a) if the policy is π_3 from above (*Answer:* 6.7);

   (b) if the policy is π_4 from above (*Answer:* 6.108).

Table 1: Estimates of state values

state	estimate
s_1	2
s_2	4
s_3	8
s_4	5
s_5	4
s_6	8
s_7	4
s_8	4
s_9	4

$V_\pi(S) = \sum_{a \in A} \pi(a|s) [R_{s,a} + \gamma V_\pi(S')]$

a) $\pi_3(S_1) = a_2$

$V_{\pi_3}(S_1) = 1.5 + 0.4 V_{\pi_3}(S_3) + 0.4 V_{\pi_3}(S_4)$.

$= 1.5 + 0.4(8) + 0.4(5)$

$= 6.7$.

b) $= 2.38 + 0.32 V_{\pi_4}(S_2) + 0.08 V_{\pi_4}(S_3) + 0.208 V_{\pi_4}(S_4) + 0.192 V_{\pi_4}(S_5)$

$= 2.38 + 0.32(4) + 0.08(8) + 0.208(5) + 0.192(4)$

$= 6.108$.

9. Consider the fragment of a gridworld in Figure 3. Let Table 1 give policy values $V_\pi(s)$. Assuming the discounting rate of $\gamma = 0.8$, calculate the values of Q-function for s_1.

Answer:

action a	$Q_\pi(s_1, a)$
a_1	5.2
a_2	6.7
a_3	6.72

$Q(S,a) = E R_{s,a} + \gamma \sum_{S' \in S} P(S'|s,a) V(S')$

$\therefore Q(S_1, a_1) = 2 + (0.8) V(S_2) = 2 + 0.8 \times 4 = 5.2$

$Q(S_1, a_2) = 0.5[1+2] + 0.8(0.5 \times 8 + 0.5 \times 5) = 6.7$

$Q(S_1, a_3) = [0.4 \times 2 + 0.6 \times 4] + 0.8[0.4 \times 5 + 0.6 \times 4] = 6.72$

10. Consider the fragment of a gridworld in Figure 3. Let Table 1 give estimates of optimal values $V^*(s)$. Update the estimate of $V^*(s_1)$ using the Bellma equation. (*Answer:* 6.72.)

$V^*(S) = \max_{a \in A} (R_{s,a} + \gamma V^*(S'))$

$\therefore V^*(S_1) = \max_{a \in A} (R_{s,a} + \gamma V^*(S'))$

$a_1 : (R_{s,a_1} + \gamma V^*(S_2)) = 5.2$

$a_2 = 6.7$

$a_3 = 6.72$.

$\therefore V^*(S_1) = Q(S_1, a_3) = 6.72$.

11. Suppose that we are evaluating the state values V_π under incomplete information settings with the discounting rate $\gamma = 0.8$. We come up with the following estimates of state values:

state	estimate of V_π
s_1	3
s_2	5

We start at state s_1 and sample an action a. It gives us reward of 2 and changes the state to s_2. What updates to the estimates can be made using TD method? Work out the update using the learning rate $\alpha = 0.1$. (*Answer:* $V_\pi(s_1)$ updates to 3.3.)

TD method: $Q_{t+1}(S,a) = Q_t(S,a) + \alpha\left([r_{s,a} + \gamma Q(S',a')] - Q_t(S,a)\right)$

$V(S) = \sum_{a \in A} \pi(a|S)\, Q_\pi(S,a)$

$S_1 \rightarrow A_1 \rightarrow R_1 \rightarrow S_2 \rightarrow A_2 \rightarrow \cdots$

t: $S_1 = 3$, $R_1 = 2$, $S_2 = 5$

$t+1: \quad Q_{t+1}(S,a) = 3 + 0.1(2 + 0.8 \times 5 - 3)$

$= 3 + 0.1(3)$

$= 3.3$ #

12. Suppose that we are evaluating the optimal values Q^* under incomplete information settings with the discounting rate $\gamma = 0.8$. At state s_1 we take action a_3, which gives us reward of 2 and takes us to state s_2. Given the following estimates of Q^*, what updates to the estimates can be made using Q-learning? Work out the update using the learning rate $\alpha = 0.1$.

action	estimate of $Q^*(s_1, a)$	estimate of $Q^*(s_2, a)$
a_1	5	4
a_2	2	2
a_3	3	5

(*Answer:* $Q^*(s_1, a_3)$ updates to 3.3.)

$Q^*(S_1, a_3) = Q_t(S_1, a_3) + \alpha\left[r_{S_1 a_3} + \gamma \max_{a \in A} Q_t(S', a') - Q_t(S_1, a_3)\right]$

$= 3 + 0.1\left[2 + 0.8(5) - 3\right]$

$= 3 + 0.1\left[2 + 4 - 3\right]$

$= 3.3$ #