MDL Assginment 2 Part I
Spring 2021 Semester

Megha Bose
2019111021

C	R
A	B

Given,

$$T(s, a, s') = P(s'|s, a) \tag{1}$$
$$P(B|A, right) = 0.8, P(A|A, right) = 0.2 \tag{2}$$
$$P(C|A, up) = 0.8, P(C|A, up) = 0.2 \tag{3}$$
$$P(A|B, left) = 0.8, P(B|B, left) = 0.2 \tag{4}$$
$$P(R|B, up) = 0.8, P(B|B, up) = 0.2 \tag{5}$$
$$P(R|C, right) = 0.25, P(C|C, right) = 0.75 \tag{6}$$
$$P(A|C, down) = 0.8, P(C|C, down) = 0.2 \tag{7}$$

---

1. **Write the Transition Table (it has to be in the form of a table).**

   Answer: The transition table is as follows:

s	a	s'	T(s, a, s')	R(s, a, s')
A	right	B	0.8	-1
A	right	A	0.2	-1
A	up	C	0.8	-1
A	up	A	0.2	-1
B	left	A	0.8	-1
B	left	B	0.2	-1
B	up	R	0.8	-4
B	up	B	0.2	-1
C	right	R	0.25	-3
C	right	C	0.75	-1
C	down	A	0.8	-1
C	down	C	0.2	-1

where s denotes start state, s' denotes successive state, a denotes action, T(s, a, s') denotes transition function and R(s, a, s') denotes reward function.

---

2. **What do you think would be the best path for the person standing at Square A to reach the Terminal State? Do not calculate anything here, just try to guess a solution through appropriate reasoning.**

Ans:
We can see that each step demands a certain cost and since we want to minimise the cost, it won't be wise to repeat states. If we look at the paths from A to R with no states repeated, we can directly see two paths,

$$A \rightarrow B \rightarrow R$$

and

$$A \rightarrow C \rightarrow R$$

.
Between these two paths we see that probability of going from A to B by taking right and probability of going from A to C by moving up have the same value of 0.8. Now, probability of going from B to R by moving up is 0.8 with cost -4 while probability of going from C to R by moving right is only 0.25 but with a cost of -3. However, the difference between the costs is just 1.
Therefore, even though path

$$A \rightarrow C \rightarrow R$$

gives better reward in total, it has huge chance, 75%, of incurring a step cost of -1 by remaining in C. We will keep incurring this cost till the agent moves to R from C.
Hence it seems better to use the path

$$A \rightarrow B \rightarrow R$$

even though it gives lesser reward.

---

3. **Perform Value Iteration by hand until convergence. Clearly draw the new values at each state after each iteration.**

Answer:

Bellman update equation for iteration t, start state s = A, B, C, successive state s' and action a:

$$U^t(s) = max_a(R(s,a) + \gamma \sum_{s'} T(s,a,s') * U^{t-1}(s'))$$

$$R(s,a) = \sum_a T(s,a,s') * R(s,a,a')$$

Therefore,

$$U^t(s) = max_a(\sum_{s'} T(s,a,s') * R(s,a,a')) + \gamma \sum_{s'} T(s,a,s') * U^{t-1}(s'))$$
$$= max_a(\sum_{s'} T(s,a,s') * (R(s,a,s') + \gamma * U^{t-1}(s')))$$

$$U^t(A) =$$
$$max(T(A,right,B) * (R(A,right,B) + \gamma * U^{t-1}(B)) + T(A,right,A) * (R(A,right,A) + \gamma * U^{t-1}(A)),$$
$$T(A,up,C) * (R(A,up,C) + \gamma * U^{t-1}(C)) + T(A,up,A) * (R(A,up,A) + \gamma * U^{t-1}(A))$$

$$= max(0.8 * (-1 + 0.2 * U^{t-1}(B)) + 0.2 * (-1 + 0.2 * U^{t-1}(A)),$$
$$0.8 * (-1 + 0.2 * U^{t-1}(C)) + 0.2 * (-1 + 0.2 * U^{t-1}(A))$$

Similarly,

$$U^t(B) = max(0.8 * (-1 + 0.2 * U^{t-1}(A)) + 0.2 * (-1 + 0.2 * U^{t-1}(B)),$$
$$0.8 * (9.9 + 0.2 * U^{t-1}(R)) + 0.2 * (-1 + 0.2 * U^{t-1}(B))$$

$$U^t(C) = max(0.25 * (10.9 + 0.2 * U^{t-1}(R)) + 0.75 * (-1 + 0.2 * U^{t-1}(C)),$$

$$0.8 * (-1 + 0.2 * U^{t-1}(A)) + 0.2 * (-1 + 0.2 * U^{t-1}(C))$$

• **Convergence:**

We iterate till the maximum change in the utility of any state in an iteration,

$$max\ diff < \frac{(1-\gamma) * error}{\gamma} = \frac{(1-0.2) * 0.01}{0.2} = 0.04$$

• **Initialisation:**

$$U^0(s) = 0$$

for s = A, B, C and

$$U^0(R) = Reward = Arr[(2019111021)\%15 = 6] = 13.9$$

.

• **Iteration t=1:**

$$U^1(A) = max(0.8 * (-1 + 0) + 0.2 * (-1 + 0) = -1, 0.8 * (-1 + 0) + 0.2 * (-1 + 0) = -1) = -1$$
$$U^1(B) = max(0.8 * (-1 + 0) + 0.2 * (-1 + 0) = -1, 0.8 * (-4 + 0.2 * 13.9) + 0.2 * (-1 + 0) = -1.176) = -1$$
$$U^1(C) = max(0.25 * (-3 + 0.2 * 13.9) + 0.75 * (-1 + 0) = -0.805, 0.8 * (-1 + 0) + 0.2 * (-1 + 0)) = -0.805$$
$$U^1(R) = 13.9$$
$$max\ diff = max(|-1-0|, |-1-0|, |-0.805-0|, |13.9-13.9|) = 1 > 0.04$$

Values are **U(A) = -1, U(B) = -1, U(C) = -0.805** and **U(R) = 13.9**.

• **Iteration t=2:**

$$U^2(A) = max(0.8 * (-1 + 0.2 * (-1)) + 0.2 * (-1 + 0.2 * (-1)) = -1.2,$$
$$0.8 * (-1 + 0.2 * (-0.805)) + 0.2 * (-1 + 0.2 * (-1)) = -1.1688) = -1.1688$$
$$U^2(B) = max(0.8 * (-1 + 0.2 * (-1)) + 0.2 * (-1 + 0.2 * (-1)) = -1.2,$$
$$0.8 * (-4 + 0.2 * 13.9) + 0.2 * (-1 + 0.2 * (-1)) = -1.216) = -1.2$$
$$U^2(C) = max(0.25 * (-3 + 0.2 * 13.9) + 0.75 * (-1 + 0.2 * (-0.805)) = -0.92575,$$
$$0.8 * (-1 + 0.2 * (-1)) + 0.2 * (-1 + 0.2 * (-0.805)) = -1.1922) = -0.92575$$
$$U^2(R) = 13.9$$
$$max\ diff = max(|-1.1688-(-1)|, |-1.2-(-1)|, |-0.92575-(-0.805)|, |13.9-13.9|) = 0.2 > 0.04$$

Values are **U(A) = -1.1688, U(B) = -1.2, U(C) = -0.92575** and **U(R) = 13.9**.

• **Iteration t=3:**

$$U^3(A) = max(0.8 * (-1 + 0.2 * (-1.2)) + 0.2 * (-1 + 0.2 * (-1.1688)) = -1.23875,$$
$$0.8 * (-1 + 0.2 * (-0.92575)) + 0.2 * (-1 + 0.2 * (-1.1688)) = -1.19487) = -1.19487$$
$$U^3(B) = max(0.8 * (-1 + 0.2 * (-1.1688)) + 0.2 * (-1 + 0.2 * (-1.2)) = -1.23501,$$
$$0.8 * (-4 + 0.2 * 13.9) + 0.2 * (-1 + 0.2 * (-1.2)) = -1.224) = -1.224$$
$$U^3(C) = max(0.25 * (-3 + 0.2 * 13.9) + 0.75 * (-1 + 0.2 * (-0.92575)) = -0.943863,$$
$$0.8 * (-1 + 0.2 * (-1.1688)) + 0.2 * (-1 + 0.2 * (-0.92575)) = -1.22404) = -0.943863$$
$$U^3(R) = 13.9$$
$$max\ diff = max(|-1.19487-(-1.1688)|, |-1.224-(-1.2)|,$$
$$|-0.943863-(-0.92575)|, |13.9-13.9|) = 0.02607 < 0.04$$

There 3 iterations are enough and final values are **U(A) = -1.19487, U(B) = -1.224, U(C) = -0.943863** and **U(R) = 13.9**.

Iteration	U(A)	U(B)	U(C)	U(R)
1	-1	-1	-0.805	13.9
2	-1.1688	-1.2	-0.92575	13.9
3	-1.19487	-1.224	-0.943863	13.9

---

4. **Find the optimal path for the person at Square A to the Terminal State using the result from Value Iteration. Was your initial guess correct?**

   Answer:

   The final values we got were **U(A) = -1.19487, U(B) = -1.224, U(C) = -0.943863** and **U(R) = 13.9**. We know that the optimal policy for state s can be given by,

$$\pi^*(s) = arg_{max,a}(\sum_{s'} P(s'|s,a)((R(s,a,s') + \gamma U(s'))$$

   . Therefore,

$$\pi^*(A) = max_a(0.8 * (-1 + 0.2 * (-1.224)) + 0.2 * (-1 + 0.2 * (-1.19487)) = -1.24363,$$

$$0.8 * (-1 + 0.2 * (-0.943863)) + 0.2 * (-1 + 0.2 * (-1.19487)) = -1.19881) = -1.19881$$

   which is for the action of

$$A \rightarrow C$$

$$\pi^*(B) = max_a(0.8 * (-1 + 0.2 * (-1.19487)) + 0.2 * (-1 + 0.2 * (-1.224)) = -1.24014,$$

$$0.8 * (-4 + 0.2 * 13.9) + 0.2 * (-1 + 0.2 * (-1.224)) = -1.22496) = -1.22496$$

   which is for the action of

$$B \rightarrow R$$

$$\pi^*(C) = max_a(0.25 * (-3 + 0.2 * 13.9) + 0.75 * (-1 + 0.2 * -0.943863)) = -0.946579,$$

$$0.8 * (-1 + 0.2 * (-1.19487)) + 0.2 * (-1 + 0.2 * (-0.943863)) = -1.22893) = -0.946579$$

   which is for the action of

$$C \rightarrow R$$

   Therefore, if a person starts at A, the optimal policy would be to go up and then right, from

$$A \rightarrow C \rightarrow R$$

   My initial guess wasn't correct. I had guessed path

$$A \rightarrow B \rightarrow R$$

   but we got the optimal policy in path

$$A \rightarrow C \rightarrow R$$

   . Though we had considered this path too in out previous analysis, I didn't guess it even after its higher reward as its probability of following the path exactly was lower than that of the other. However, we see after performing value iteration algorithm that path

$$A \rightarrow C \rightarrow R$$

   is preferred. It might be because the path gives higher reward and the agent aims to collect maximum reward it can.

---

5. **Try to make a guess of what could be the importance of specific values of reward and what could be the possible trend with the different values of the reward.**

Answer:

We see that the agent focuses on increasing its reward and chooses the path that maximises the total reward. From path

$$A \rightarrow B \rightarrow R$$

, we get reward of -1-3+13.9 = 9.9 while from path

$$A \rightarrow C \rightarrow R$$

, we get reward of -1-4+13.9 = 8.9. We saw after applying value iteration algorithm that path

$$A \rightarrow C \rightarrow R$$

was preferred.
However, if we increase the utility at terminal state (R), the agent might start taking the policy with lower risk of incurring more losses.
When R value is low, the difference between rewards got from the two paths, which is 1, may be a significant value. Since the agent won't want to lose the reward, it chooses the risky but rewarding path. However, when the value of R increases, this difference of 1 between the rewards got from the two paths becomes less and less significant. In this case, the agent may try to focus more on the huge difference between the probabilities of reaching R from B (0.8) and R from C (0.25). The reward factor is dominated by this risk factor and hence path

$$A \rightarrow B \rightarrow R$$

might be preferred as R increases beyond a certain value. Thus we can say that the optimality of the path

$$A \rightarrow C \rightarrow R$$

is more for smaller values of R while for larger values of R, optimality of

$$A \rightarrow B \rightarrow R$$

is more.

*** (O – O) ***

P.S. Moodle asks for handwritten document but Jai Bardhan Sir (TA) told it was okay to write the assignment in LaTeX.