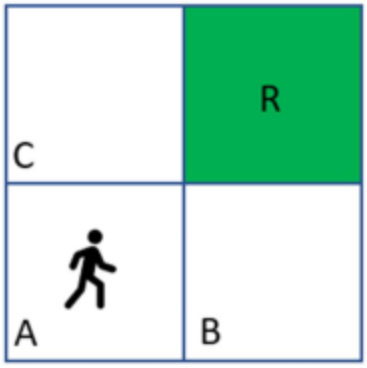
MDL Assignment-2

Aakash Aanegola
2019101009

Transition table

Current state	Action	Probability	Next state	Step cost
A	Move up	0.8	C	-1
A	Move up	0.2	A	-1
A	Move right	0.8	B	-1
A	Move right	0.2	A	-1
B	Move up	0.8	R	-4
B	Move up	0.2	B	-1
B	Move left	0.8	A	-1
B	Move left	0.2	B	-1
C	Move down	0.8	A	-1
C	Move down	0.2	C	-1
C	Move right	0.25	R	-3
C	Move right	0.75	C	-1

In the above table, $\{A, B, C\}$ represent the states, and R represents the final state. The reward at R = 20



The gridworld for this problem

What I guess the optimal policy (path) is

As we can observe from the grid above, the only viable paths are $A \rightarrow B \rightarrow R$ and $A \rightarrow C \rightarrow R$. This is because each step incurs a penalty of -1 (step cost) and hence moving from C or B back to A would only serve to reset our position after incurring additional cost.

Now that we know that our agent will follow one of two paths, we can observe that traversing the pathway $A \rightarrow B$ is equivalent to the pathway $A \rightarrow C$. This is because the probabilities of successful transitions and the step cost for all scenarios are the same. The only difference between these paths is between the pathways $B \rightarrow R$ and $C \rightarrow R$.

We know that the discount factor $\gamma = 0.2$ which means that the rewards our agent gains at the next step is worth 0.2 times the reward if it was gained at this step. Seeing as the probability of successful transition is 0.8 from B and the step cost is -4 compared to the probability of successful transition of 0.25 from C with step cost -3 intuitively the pathway that traverses through B is more optimal than the pathway that passes through C (the probability that the agent spends longer at C incurring step costs waiting to escape the state is higher than the probability of staying at state B. Additionally if the agent stays at state C for 2 steps longer than it stays at state B, it incurs a higher step cost which is very probable). However since the agent factors in the discount factor, there may be a threshold below which the pathway through C is more optimal as the immediate reward is more than that at B and the final reward may not be enough to compensate for this additional current cost.

Given that the final reward R is equal to 20, I believe it is safe to say that the reward is high enough to account for the higher step cost of the pathway through B. Therefore, my guess for the optimal path is $A \rightarrow B \rightarrow R$, or expressed differently:

Optimal policy (actions to be taken at each state)

Current state	Action
A	Move right
B	Move up
C	Move right

The value iteration algorithm

The following notation has been used for the value iteration diagram.

C	R
A	B

Where A, B, C and R represent the value at their respective states.

Setup:

The grid was initialized with the values :

$A = 0$

$B = 0$

$C = 0$

$R = 20$

This is because initially we do not know the values of each of the states (apart from the final state). The final state has a value of 20 that does not change, because once the agent reaches the final state it does not attempt to take any more actions.

The bellman equation:

The bellman equation is as follows

$$V_{t+1}(s) = \max_a \sum_{s'} P(s, a, s^{'})[R(s, a, s^{'}) + \gamma V_t(s^{'})]$$

where,

$V_t(s)$ represents the value (utility) of a state at time t,

$R(s, a, s^{'})$ represents the step cost of taking an action a from state s and ending up in state $s^{'}$,

$P(s, a, s^{'})$ represents the probability of transitioning to state $s^{'}$ from state s given that the agent takes action a.

As we know, the value iteration algorithm uses this formula and takes the best action (quantified by the $\max$ operation over all the rewards for taking a particular action) iteratively until the maximum change in the value of any state (δ) is less than a certain threshold, $\Delta = 0.01$ in our case.

To illustrate the value iteration algorithm, let us take a look at one iteration for state A.

$$V_{t+1}(A) = \max(0.8[-1 + \gamma V_t(B)] + 0.2[-1 + \gamma V_t(A)] \, , \\ 0.8[-1 + \gamma V_t(C)] + 0.2[-1 + \gamma V_t(A)])$$

The equations for states B and C are very similar, differing only in the probabilities of transitions, states and the step cost (reward). As is easily observable, the reward at state R will remain constant as there is no action that can be taken from it.

Initial values:

0	20
0	0

Iteration 1:

$\delta = 1$

-0.5	20
-1	-0.2

Iteration 2:

$\delta = 0.075$

-0.575	20
-1.072	-0.208

Iteration 3:

$\delta = 0.011$

-0.586	20
-1.076	-0.208

Iteration 4:

$\delta = 0.002$

-0.588	20
-1.076	-0.208

Optimal path based on value iteration

The optimal path for the agent is the path that gets the agent maximal reward. Since the value (utility) of state B is more than that of state C the agent will prefer moving along the path $A \rightarrow B \rightarrow R$.

The guess I made for the policy as well as the path were correct.

Reward value vs optimal policy

Let us take a look at what the value of the discount factor entails. If $\gamma = 0$, then our agent plays greedily for only the current action. This means that the agent will try to maximize it's reward at each step and disregard any future rewards this means that the path $A \rightarrow C \rightarrow R$ will be optimal regardless of the final reward. On the other hand, if $\gamma = 1$, the agent cares about the total reward and hence will always prefer the path $A \rightarrow B \rightarrow R$. Since the discount factor in this case is 0.2, we will have a tipping point where the agent 'switches' from preferring the path through C to the path through B.

Additionally, the optimal policy depends on the step cost. If the step cost is low then the agent can spend time trying to maximize the reward, however if the step cost is high then the agent will try to get to the final state irrespective of the cost (making more moves means a higher penalty).
In the case of our gridworld, it is very likely that there is a tipping point for the reward. What this means is that if the reward is below a certain threshold, R_0 for instance the agent will prefer the pathway $A \rightarrow C \rightarrow R$. However if the final reward is above the final threshold R_0 then the agent will start preferring the

pathway $A \rightarrow B \rightarrow R$. This is due to the discount factor, and the fact that the agent prefers gaining a reward now over in the next step. The actions at state C have a higher reward than those at state B, and the expected reward is also higher in state C. This means that if the final reward is low, the agent will prefer getting a higher reward in the current state (namely C) over the low reward it would get if it was at state B. However if the reward is high enough, then the pathway through B is preferred as the higher step cost is compensated by the higher transition success probability and the less time making unwanted steps.