MDL Assignment 2

Part 1

- Rutvij Menavlikar
(20191111032)

1. The transition tables for actions {Up, Right, Down, Left} are as follows. (Let the terminal state be R, and states A, B, C as given in problem statement)

Up

	A	B	C	R
A	0.2	0	0.8	0
B	0	0.2	0	0.8
C	0	0	0	0
R	0	0	0	0

Right

	A	B	C	R
A	0.2	0.8	0	0
B	0	0	0	0
C	0	0	0.75	0.25
R	0	0	0	0

Down

	A	B	C	R
A	0	0	0	0
B	0	0	0	0
C	0.8	0	0.2	0
R	0	0	0	0

Left

	A	B	C	R
A	0	0	0	0
B	0.8	0.2	0	0
C	0	0	0	0
R	0	0	0	0

2. Any valid path which is not one of A-B-R and A-C-R would include going from B to C or C to B via A which will unnecessarily increase step cost.
The probability and cost for going from A to B and A to C are the same. Therefore, the 2 paths will be compared by the expected cost of going from B to R and C to R.
The probability of going from B to R in one move is much more than that of going from C to R, but the step cost of the latter is less than that of the former.
Finding the expected values for the transition in these states, we need to calculate summations of Arithmetico-Geometric Progressions. But, to get a rough estimate just by looking at the (cost, probability) pairs (B→R: (4, 0.8), C→R: (3, 0.25)) a rough estimate can be made that the expected cost of going from B to R is less

because the difference in expected number of moves exceeds the difference between initial step cost, due to high probability of going B→R in one move. Hence, $A \to B \to R$ is estimated to be the best path.

3. Reward function for state-action pairs $R(s,a)$ is as follows

$R(A, \text{"Up"}) = (-1)0.8 + (-1)0.2 = -1$ $\quad R(A, \text{"Right"}) = (-1)0.8 + (-1)0.2 = -1$

$R(B, \text{"Left"}) = (-1)0.8 + (-1)0.2 = -1$ $\quad R(B, \text{"Up"}) = (-4)0.8 + (-1)0.2 = -3.4$

$R(C, \text{"Down"}) = (-1)0.8 + (-1)0.2 = -1$ $\quad R(C, \text{"Right"}) = (-3)0.25 + (-1)0.75 = -1.5$

For all remaining state-action pairs, $R(s,a) = 0$ as those actions are invalid for the corresponding states.

We will consider the initial utility function as the gained reward for each state. That is, $U_0 = \{A:0, B:0, C:0, R:10\}$

For each iteration, we use the Bellman Equation:

$U'(s) = \max_{a \in A(s)} \{ R(s,a) + \gamma \sum_{s'} U(s') P(s'|s,a) \}$ and then assign values of U' to U.

First Iteration:

$U_1(A) = \max \{ R(A, \text{"Right"}) + \gamma U_0(B) P(B|A, \text{"Right"}) + \gamma U_0(A) P(A|A, \text{"Right"}),$
$R(A, \text{"Up"}) + \gamma U_0(C) P(C|A, \text{"Up"}) + \gamma U_0(A) P(A|A, \text{"Up"}) \}$

$= \max \{ -1 + (0.2)(0)(0.8) + (0.2)(0)(0.2), -1 + (0.2)(0)(0.8) + (0.2)(0)(0.2) \}$

$= -1$

$U_1(B) = \max \{ R(B, \text{"Left"}) + \gamma U_0(A) P(A|B, \text{"Left"}) + \gamma U_0(B) P(B|B, \text{"Left"}),$
$R(B, \text{"Up"}) + \gamma U_0(R) P(R|B, \text{"Up"}) + \gamma U_0(B) P(B|B, \text{"Up"}) \}$

$= \max \{ -1 + (0.2)(0)(0.8) + (0.2)(0)(0.2), -3.4 + (0.2)(10)(0.8) + (0.2)(0)(0.2) \}$

$= -1$

$U_1(C) = \max \{ R(C, \text{"Down"}) + \gamma U_0(A) P(A|C, \text{"Down"}) + \gamma U_0(C) P(C|C, \text{"Down"}),$
$R(C, \text{"Right"}) + \gamma U_0(R) P(R|C, \text{"Right"}) + \gamma U_0(C) P(C|C, \text{"Right"}) \}$

$= \max \{ -1 + (0.2)(0)(0.8) + (0.2)(0)(0.2), -1.5 + (0.2)(10)(0.25) + (0.2)(0)(0.75) \}$

$= -1$

$\therefore U_1 = \{A:-1, B:-1, C:-1, R:10\}$, max. deviation $= 1$ $\quad (> 0.01)$

Second Iteration:

$$U_2(A) = \max\{R(A,\text{"Right"}) + \gamma U_1(B)P(B|A,\text{"Right"}) + \gamma U_1(A)P(A|A,\text{"Right"}),\ R(A,\text{"Up"}) + \gamma U_1(C)P(C|A,\text{"Up"}) + \gamma U_1(A)P(A|A,\text{"Up"})\}$$
$$= \max\{-1 + (0.2)(-1)(0.8) + (0.2)(-1)(0.2),\ -1 + (0.2)(-1)(0.8) + (0.2)(-1)(0.2)\}$$
$$= -1.2$$

$$U_2(B) = \max\{R(B,\text{"Left"}) + \gamma U_1(A)P(A|B,\text{"Left"}) + \gamma U_1(B)P(B|B,\text{"Left"}),\ R(B,\text{"Up"}) + \gamma U_1(R)P(R|B,\text{"Up"}) + \gamma U_1(B)P(B|B,\text{"Up"})\}$$
$$= \max\{-1 + (0.2)(-1)(0.8) + (0.2)(-1)(0.2),\ -3.4 + (0.2)(10)(0.8) + (0.2)(-1)(0.2)\}$$
$$= -1.2$$

$$U_2(C) = \max\{R(C,\text{"Down"}) + \gamma U_1(A)P(A|C,\text{"Down"}) + \gamma U_1(C)P(C|C,\text{"Down"}),\ R(C,\text{"Right"}) + \gamma U_1(R)P(R|C,\text{"Right"}) + \gamma U_1(C)P(C|C,\text{"Right"})\}$$
$$= \max\{-1 + (0.2)(-1)(0.8) + (0.2)(-1)(0.2),\ -1.5 + (0.2)(10)(0.25) + (0.2)(-1)(0.75)\}$$
$$= -1.15$$

$\therefore U_2 = \{A:-1.2,\ B:-1.2,\ C:-1.15,\ R:10\}$, max. deviation $= 0.2$ (>0.01)

Third Iteration:

$$U_3(A) = \max\{R(A,\text{"Right"}) + \gamma U_2(B)P(B|A,\text{"Right"}) + \gamma U_2(A)P(A|A,\text{"Right"}),\ R(A,\text{"Up"}) + \gamma U_2(C)P(C|A,\text{"Up"}) + \gamma U_2(A)P(A|A,\text{"Up"})\}$$
$$= \max\{-1 + (0.2)(-1.2)(0.8) + (0.2)(-1.2)(0.2),\ -1 + (0.2)(-1.15)(0.8) + (0.2)(-1.2)(0.2)\}$$
$$= -1.232$$

$$U_3(B) = \max\{R(B,\text{"Left"}) + \gamma U_2(A)P(A|B,\text{"Left"}) + \gamma U_2(B)P(B|B,\text{"Left"}),\ R(B,\text{"Up"}) + \gamma U_2(R)P(R|B,\text{"Up"}) + \gamma U_2(B)P(B|B,\text{"Up"})\}$$
$$= \max\{-1 + (0.2)(-1.2)(0.8) + (0.2)(-1.2)(0.2),\ -3.4 + (0.2)(10)(0.8) + (0.2)(-1.2)(0.2)\}$$
$$= -1.24$$

$$U_3(C) = \max\{R(C,\text{"Down"}) + \gamma U_2(A)P(A|C,\text{"Down"}) + \gamma U_2(C)P(C|C,\text{"Down"}),\ R(C,\text{"Right"}) + \gamma U_2(R)P(R|C,\text{"Right"}) + \gamma U_2(C)P(C|C,\text{"Right"})\}$$
$$= \max\{-1 + (0.2)(-1.2)(0.8) + (0.2)(-1.15)(0.2),\ -1.5 + (0.2)(10)(0.25) + (0.2)(-1.15)(0.75)\}$$
$$= -1.1725$$

$\therefore U_3 = \{A:-1.232,\ B:-1.24\ C:-1.1725,\ R:10\}$, max. deviation $= 0.04$ (>0.01)

Fourth Iteration:

$U_4(A) = \max\{R(A,\text{"Right"}) + \gamma U_3(B)P(B|A,\text{"Right"}) + \gamma U_3(A)P(A|A,\text{"Right"}),$
$R(A,\text{"Up"}) + \gamma U_3(C)P(C|A,\text{"Up"}) + \gamma U_3(A)P(A|A,\text{"Up"})\}$
$= \max\{-1 + (0.2)(-1.24)(0.8) + (0.2)(-1.232)(0.2), -1 + (0.2)(-1.1725)(0.8) + (0.2)(-1.232)(0.2)\}$
$= -1.23688$

$U_4(B) = \max\{R(B,\text{"Left"}) + \gamma U_3(A)P(A|B,\text{"Left"}) + \gamma U_3(B)P(B|B,\text{"Left"}),$
$R(B,\text{"Up"}) + \gamma U_3(R)P(R|B,\text{"Up"}) + \gamma U_3(B)P(B|B,\text{"Up"})\}$
$= \max\{-1 + (0.2)(-1.232)(0.8) + (0.2)(-1.24)(0.2), -3.4 + (0.2)(10)(0.8) + (0.2)(-1.24)(0.2)\}$
$= -1.24672$

$U_4(C) = \max\{R(C,\text{"Down"}) + \gamma U_3(A)P(A|C,\text{"Down"}) + \gamma U_3(C)P(C|C,\text{"Down"}),$
$R(C,\text{"Right"}) + \gamma U_3(R)P(R|C,\text{"Right"}) + \gamma U_3(C)P(C|C,\text{"Right"})\}$
$= \max\{-1 + (0.2)(-1.232)(0.8) + (0.2)(-1.1725)(0.2), -1.5 + (0.2)(10)(0.25) + (0.2)(-1.1725)(0.75)\}$
$= -1.175875$

$\therefore U_4 = \{A:-1.23688, B:-1.24672, C:-1.175875, R:10\}$, max. deviation $= 0.00672$ (≤ 0.01)
The Utility function has now converged, according to given constraints.

4. To find the optimal path from square A to terminal state, we use the utility function calculated by value-iteration.
That is for every state s, we choose an action $a' \in A(s)$ such that
$R(s,a') + \gamma \sum_{s' \in S} U(s')P(s'|s,a') = \max_{a \in A(s)} \{R(s,a) + \gamma \sum_{s' \in S} U(s')P(s'|s,a)\}$, i.e. we choose the action that would maximise utility function value if we were to perform another iteration. Our base case is A.
$U(A) = -1.23688$, $U(B) = -1.24672$, $U(C) = -1.175875$, $U(R) = 10$

A:

"Right": $R(A,\text{"Right"}) + \gamma U(B)P(B|A,\text{"Right"}) + \gamma U(A)P(A|A,\text{"Right"})$
$= -1 + (0.2)(-1.24672)(0.8) + (0.2)(-1.23688)(0.2) = -1.2489504$

"Up": $R(A,\text{"Up"}) + \gamma U(C)P(C|A,\text{"Up"}) + \gamma U(A)P(A|A,\text{"Up"})$
$= -1 + (0.2)(-1.175875)(0.8) + (0.2)(-1.23688)(0.2) = -1.2376152$

∴ The optimal move from state A is to move "Up", i.e. go to square C.

C:

"Down": $R(C, \text{"Down"}) + \gamma U(A) P(A|C, \text{"Down"}) + \gamma U(C) P(C|C, \text{"Down"})$

$= -1 + (0.2)(-1.23688)(0.8) + (0.2)(-1.175875)(0.2) = -1.2449358$

"Right": $R(C, \text{"Right"}) + \gamma U(R) P(R|C, \text{"Right"}) + \gamma U(C) P(C|C, \text{"Right"})$

$= -1.5 + (0.2)(10)(0.25) + (0.2)(-1.175875)(0.75) = -1.17638125$

∴ The optimal move from state C is to move "Right", i.e. go to terminal state.

Hence, the optimal path from square A to the terminal state is A → C → R. And, my initial guess was not correct.

5. As explained in Q2, the comparison of the 2 paths A-B-R and A-C-R is reduced to comparison between transitions B→R and C→R. Cost of B→R is higher than C→R, but so is the probability of transition.

With a higher reward, the effect of B→R transition's higher cost on choice of optimal path reduces, and the effect of B→R transition's higher probability increases.

Hence, after crossing a certain value of reward, the optimal path changes to A→B→R from A→C→R.