MDL Assignment 3

ARJUN DOSAJH
2021113016

10	Reward: 1 11	Penalty: -1 12
7	8	9
4	Wall 5	6
Start 1	2	3

$$V_0(1) = V_0(2) = V_0(3) = V_0(4) = V_0(6) = V_0(7) = V_0(8) = V_0(9) = V_0(10) = 0$$

$V_0(11) = +1 \longrightarrow$ Reward

$V_0(12) = -1 \longrightarrow$ Penalty

Iteration 1

$$V_1(I) = \max_A \left[R(I,A) + \gamma \sum_J P(J|I,A) * V_0(J) \right]$$

since γ and $R(I,A)$ are constant always, we can write

$$V_1(I) = R(I,A) + \gamma \max_A \left(\sum P(J|I,A) * V_0(J) \right)$$

$$V_1(I) = -0.04 + 0.95 \max_A \left(\sum P(J|I,A) * V_0(J) \right)$$

$$V_1(1) = -0.04 + 0.95 \max \begin{pmatrix} \sum P(J \mid I, A = \text{"UP"}) * V_0(J) \\ \sum P(J \mid I, A = \text{"DOWN"}) * V_0(J) \\ \sum P(J \mid I, A = \text{"LEFT"}) * V_0(J) \\ \sum P(J \mid I, A = \text{"RIGHT"}) * V_0(J) \end{pmatrix}$$

$$= -0.04 + 0.95 \max \begin{pmatrix} 0.7\,V_0(4) + 0.15\,V_0(1) + 0.15\,V_0(2) \\ 0.7\,V_0(1) + 0.15\,V_0(1) + 0.15\,V_0(2) \\ 0.7\,V_0(1) + 0.15\,V_0(1) + 0.15\,V_0(4) \\ 0.7\,V_0(2) + 0.15\,V_0(1) + 0.15\,V_0(4) \end{pmatrix}$$

$$= -0.04 + 0.95 \max \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = -0.04$$

$\therefore$ $V_1(1) = -0.04$

$V_1(2) = -0.04$

$V_1(3) = -0.04$

$V_1(4) = -0.04$

$V_1(6) = -0.04$

$V_1(7) = -0.04$

$V_1(9) = -0.04$

$$V_1(10) = -0.04 + 0.95 \max \begin{pmatrix} 0.7\,V_0(10) + 0.15\,V_0(10) + 0.15\,V_0(11) \\ 0.7\,V_0(7) + 0.15\,V_0(10) + 0.15\,V_0(11) \\ 0.7\,V_0(10) + 0.15\,V_0(10) + 0.15\,V_0(7) \\ 0.7\,V_0(11) + 0.15\,V_0(10) + 0.15\,V_0(7) \end{pmatrix}$$

$$= -0.04 + 0.95 \max \begin{pmatrix} 0.7(0) + 0.15(0) + 0.15(1) \\ 0.7(0) + 0.15(0) + 0.15(1) \\ 0.7(0) + 0.15(0) + 0.15(0) \\ 0.7(1) + 0.15(0) + 0.15(0) \end{pmatrix}$$

$$= -0.04 + 0.95 \begin{pmatrix} 0.15 \\ 0.15 \\ 0 \\ 0.7 \end{pmatrix} = -0.04 + (0.95)(0.7)$$

$$V_1(10) = 0.625$$

Similarly, $V_1(8) = 0.625$

$\therefore$ new_utilities = [-0.04 (cell 1), -0.04 (cell 2), -0.04, -0.04, 0 (wall), -0.04,
-0.04, 0.625, -0.04, 0.625, 1, -1]

0.625 → $V_1(8)$; 0.625 → $V_1(10)$; 1 → reward; -1 → penalty; cell 12

We get the same result from the code after 1 iteration

```
Iteration 1: -0.04 -0.04 -0.04 -0.04 0 -0.04 -0.04 0.625 -0.04 0.625 1 -1
```

Iteration 2

We can directly say that

$V_2(1) = V_2(2) = V_2(3) = V_2(4) = V_2(6)$, since all the adjacent states to these cells have the same utility value $= -0.04$

$$V_2(1) = -0.04 + 0.95 \max \begin{pmatrix} 0.7V_1(4) + 0.15V_1(1) + 0.15V_1(2) \\ 0.7V_1(1) + 0.15V_1(1) + 0.15V_1(2) \\ 0.7V_1(1) + 0.15V_1(4) + 0.15V_1(1) \\ 0.7V_1(2) + 0.15V_1(1) + 0.15V_1(4) \end{pmatrix}$$

$$= -0.04 + 0.95 \max \begin{pmatrix} (-0.04)(0.7 + 0.15 + 0.15) \\ (-0.04)(0.7 + 0.15 + 0.15) \\ (-0.04)(0.7 + 0.15 + 0.15) \\ (-0.04)(0.7 + 0.15 + 0.15) \end{pmatrix}$$

$$= -0.04 + 0.95 \begin{pmatrix} -0.04 \\ -0.04 \\ -0.04 \\ -0.04 \end{pmatrix} = -0.04 + (-0.04)(0.95)$$

$$= -0.078$$

$\therefore V_2(1) = V_2(2) = V_2(3) = V_2(4) = V_2(6) = -0.078$

$$V_2(7) = -0.04 + 0.95 \max \begin{pmatrix} 0.7V_1(10) + 0.15V_1(7) + 0.15V_1(8) \\ 0.7V_1(4) + 0.15V_1(8) + 0.15V_1(7) \\ 0.7V_1(7) + 0.15V_1(10) + 0.15V_1(4) \\ 0.7V_1(8) + 0.15V_1(10) + 0.15V_1(4) \end{pmatrix}$$

$$= -0.04 + 0.95 \max \begin{pmatrix} 0.525 \\ 0.059 \\ 0.059 \\ 0.525 \end{pmatrix}$$

$$= -0.04 + (0.95 \times 0.525)$$

$$= 0.459$$

$$V_2(8) = -0.04 + 0.95 \max \begin{pmatrix} 0.7V_1(11) + 0.15V_1(7) + 0.15V_1(9) \\ 0.7V_1(8) + 0.15V_1(7) + 0.15V_1(9) \\ 0.7V_1(7) + 0.15V_1(8) + 0.15V_1(11) \\ 0.7V_1(9) + 0.15V_1(8) + 0.15V_1(11) \end{pmatrix}$$

$$= -0.04 + 0.95 \max \begin{pmatrix} 0.688 \\ 0.425 \\ 0.215 \\ 0.215 \end{pmatrix} = -0.04 + 0.95(0.688) = 0.614$$

$$V_2(9) = -0.04 + 0.95 \max \begin{pmatrix} 0.7V_1(12) + 0.15V_1(8) + 0.15V_1(9) \\ 0.7V_1(6) + 0.15V_1(8) + 0.15V_1(9) \\ 0.7V_1(8) + 0.15V_1(12) + 0.15V_1(6) \\ 0.7V_1(9) + 0.15V_1(12) + 0.15V_1(6) \end{pmatrix}$$

$$= -0.04 + 0.95 \max \begin{pmatrix} -0.612 \\ 0.059 \\ 0.281 \\ -0.184 \end{pmatrix} = -0.04 + (0.95 \times 0.281)$$

$$= 0.227$$

$$V_2(10) = -0.04 + 0.95 \max \begin{pmatrix} 0.7\,V_1(10) + 0.15\,V_1(10) + 0.15\,V_1(11) \\ 0.7\,V_1(7) + 0.15\,V_1(11) + 0.15\,V_1(10) \\ 0.7\,V_1(10) + 0.15\,V_1(7) + 0.15\,V_1(10) \\ 0.7\,V_1(11) + 0.15\,V_1(7) + 0.15\,V_1(10) \end{pmatrix}$$

$$= -0.04 + 0.95 \max \begin{pmatrix} 0.681 \\ 0.215 \\ 0.525 \\ 0.787 \end{pmatrix} = -0.04 + (0.95 \times 0.787)$$

$$= 0.708$$

$\therefore$ new utilities $= [-0.078, -0.078, -0.078, -0.078, 0, -0.078,$
$0.459, 0.614, 0.227, 0.708, 1, -1]$

We get the same result from the code after 2 iterations

```
Iteration 2: -0.078 -0.078 -0.078 -0.078 0 -0.078 0.459 0.614 0.227 0.708 1 -1
```